Shallow Learning Tends to Ridgelet Transform

Sho Sonoda

September 27, 2026

Chapter 1

Introduction

This blueprint is generated from the `@[blueprint]` annotations in the Lean sources. Edit the Lean files, then run `./script/generate.sh`.

Chapter 2

Setting

The setting of the paper (Section 2): feature maps, the tanh feature of assumption (A3), particle measures and their realization, hidden marginal, conditional mean amplitude and coefficient measure, the synthesis and analysis operators, the kernel operator, its resolvent and the regularized ridgelet transform, the risks, the Gibbs reference measure, the domain and the free energy, and the finite-width and finite-sample objects.

2.1 Feature maps

Definition 1. Let $\mathcal{X}$ and Z be measurable spaces. A *feature map* is a jointly measurable function $\varphi\colon Z \times \mathcal{X} \to \mathbb{R}$, $(z, x) \mapsto \varphi_z(x)$, with $|\varphi_z(x)| \le 1$ for all $z \in Z$ and $x \in \mathcal{X}$. Assumption (A3) of the paper, $\varphi_{(w,b)}(x) = \tanh(w^\top x - b)$ on $Z = \mathbb{R}^m \times \mathbb{R}$, is the instance 'tanhFeature'.

Definition 2. Let Z be a metric space. The feature map φ is *Lipschitz in the hidden parameter* if there is $L \ge 0$ such that $|\varphi_z(x) - \varphi_{z'}(x)| \le L\, d(z, z')$ for all $x \in \mathcal{X}$ and $z, z' \in Z$. Under (A1) and (A3) this holds with $L = \sqrt{R_X^2 + 1}$.

Definition 3. For a feature map φ on $Z \times \mathcal{X}$ and a subset $A \subseteq \mathcal{X}$, the *restriction* of φ to A is the feature map $Z \times A \to \mathbb{R}$, $(z, x) \mapsto \varphi_z(x)$; it is used with $A = \{|x| \le R_X\}$, which is assumption (A1).

Definition 4. For a feature map φ on $Z \times \mathcal{X}$ and a measurable map $e\colon Z' \to Z$, the *reparametrized* feature map on $Z' \times \mathcal{X}$ is $(z', x) \mapsto \varphi_{e(z')}(x)$. It is used to identify the hidden-parameter space $\mathbb{R}^m \times \mathbb{R}$ of (A3) with the Euclidean space $\mathbb{R}^{m+1}$.

Definition 5. (A3) Let $m \ge 1$, $\mathcal{X} = \mathbb{R}^m$ and $Z = \mathbb{R}^m \times \mathbb{R}$. The hidden parameter is $z = (w, b)$ and $\varphi_z(x) := \tanh(w^\top x - b)$. Then $|\varphi_z| \le 1$ and $(z, x) \mapsto \varphi_z(x)$ is continuous, so φ is a feature map in the sense of Definition 1.

Definition 6. (A3) with the Euclidean hidden-parameter space: the tanh feature $\varphi_{(w,b)}(x) = \tanh(w^\top x - b)$ of Definition 5 with $z = (w, b)$ in $\mathbb{R}^{m+1} = \mathbb{R}^m \times \mathbb{R}$ carrying the Euclidean norm $|(w, b)| = \sqrt{|w|^2 + b^2}$ (the reparametrization, Definition 4, along the identity $\mathbb{R}^{m+1} \to \mathbb{R}^m \times \mathbb{R}$).

2.2 Particle measures

Definition 7. For $\rho \in \mathcal{P}(\Theta)$, $\Theta = \mathbb{R} \times Z$, with $\int |a|\, \mathrm{d}\rho < \infty$ the *realization* is $F_\rho(x) := \int_\Theta a\, \varphi_z(x)\, \rho(\,\mathrm{d}a\, \mathrm{d}z)$, so that $|F_\rho(x)| \le \int |a|\, \mathrm{d}\rho$.

Definition 8. The *hidden marginal* of $\rho \in \mathcal{P}(\Theta)$ is $\nu_\rho(\mathrm{d}z) := \int_{\mathbb{R}} \rho(\mathrm{d}a\,\mathrm{d}z)$, the image of ρ under $(a, z) \mapsto z$.

Definition 9. The *conditional mean amplitude* of $\rho \in \mathcal{P}(\Theta)$ is $m_\rho(z) := \mathbb{E}_\rho[a \mid z]$, defined ν_ρ-a.e. through the disintegration $\rho(\mathrm{d}a\,\mathrm{d}z) = \nu_\rho(\mathrm{d}z)\,\kappa(z, \mathrm{d}a)$ as $m_\rho(z) = \int_{\mathbb{R}} a\,\kappa(z, \mathrm{d}a)$.

Definition 10. The *coefficient measure* of $\rho \in \mathcal{P}(\Theta)$ with $\int |a|\,\mathrm{d}\rho < \infty$ is the signed measure $\Pi\rho(\mathrm{d}z) := \int_{\mathbb{R}} a\,\rho(\mathrm{d}a\,\mathrm{d}z)$ on Z, i.e. $\Pi\rho(A) = \int_{\mathbb{R}\times A} a\,\mathrm{d}\rho$; by disintegration $\Pi\rho = m_\rho\,\nu_\rho$.

Definition 11. For particles $\theta = (\theta_\ell)_{\ell=1}^M \in \Theta^M$ the *empirical particle measure* is $\rho_\theta := \frac{1}{M}\sum_{\ell=1}^M \delta_{\theta_\ell} \in \mathcal{P}(\Theta)$.

Theorem 12. *If $\int |a|\,\mathrm{d}\rho < \infty$ then, for every $x \in \mathcal{X}$, $\theta \mapsto a\,\varphi_z(x)$ is ρ-integrable, since $|a\varphi_z(x)| \le |a|$.*

Proof. □

Theorem 13. *If $\int |a|\,\mathrm{d}\rho < \infty$ then the realization $F_\rho(x) = \int_\Theta a\,\varphi_z(x)\,\rho(\mathrm{d}\theta)$ is a well-defined bounded function with $|F_\rho(x)| \le \int |a|\,\mathrm{d}\rho$ for every $x \in \mathcal{X}$.*

Proof. □

2.3 Synthesis, analysis, kernel operator and regularized ridgelet transform

Definition 14. For $r \in L^2(P_X)$ the *analysis map* is $(S^*r)(z) := \langle \varphi_z, r\rangle_{L^2(P_X)} = \int_{\mathcal{X}} \varphi_z(x) r(x)\,P_X(\mathrm{d}x)$, a function of $z \in Z$ that does not depend on the base measure ν. By Lemma **??** it is the adjoint of S_ν for every ν.

Definition 15. For a base measure $\nu \in \mathcal{P}(Z)$ and $u\colon Z \to \mathbb{R}$ the *synthesis* of u is the function $(S_\nu u)(x) := \int_Z u(z)\varphi_z(x)\,\nu(\mathrm{d}z)$ on $\mathcal{X}$.

Definition 16. For $\nu \in \mathcal{P}(Z)$ the *synthesis operator* is $S_\nu\colon L^2(\nu) \to L^2(P_X)$, $(S_\nu u)(x) := \int_Z u(z)\varphi_z(x)\,\nu(\mathrm{d}z)$. It is a well-defined bounded linear operator with $\|S_\nu\| \le 1$, since $|(S_\nu u)(x)| \le \int |u|\,\mathrm{d}\nu \le \|u\|_{L^2(\nu)}$ pointwise.

Definition 17. The adjoint $S_\nu^*\colon L^2(P_X) \to L^2(\nu)$ of the synthesis operator. By Lemma **??**, $(S_\nu^* r)(z) = \langle \varphi_z, r\rangle_{L^2(P_X)} = (S^*r)(z)$ ν-a.e.

Definition 18. The *kernel operator* on $L^2(P_X)$ is $K_\nu := S_\nu S_\nu^*$; its integral kernel is $k_\nu(x, x') = \int \varphi_z(x)\varphi_z(x')\,\nu(\mathrm{d}z)$.

Definition 19. For $\lambda > 0$ the *resolvent* $(K_\nu + \lambda)^{-1}$ is the inverse of the bounded operator $K_\nu + \lambda$ on $L^2(P_X)$; by Lemma **??** it exists and $\|(K_\nu + \lambda)^{-1}\| \le 1/\lambda$. (In Lean it is defined as the inverse when $K_\nu + \lambda$ is invertible and as 0 otherwise.)

Definition 20. For $\lambda > 0$, $\nu \in \mathcal{P}(Z)$ and $f \in L^2(P_X)$ the *regularized ridgelet transform* is $R_{\lambda,\nu} f := S^*(K_\nu + \lambda)^{-1} f$, i.e. $(R_{\lambda,\nu} f)(z) = \langle \varphi_z, (K_\nu + \lambda)^{-1} f\rangle_{L^2(P_X)}$, a bounded continuous function of z that lies in $L^2(\nu)$.

2.4 Risks, Gibbs reference measure and free energy

Definition 21. For $f \in L^2(P_X)$ the *population risk* of $\rho \in \mathcal{P}(\Theta)$ is $L(\rho) := \frac{1}{2}\|F_\rho - f\|^2_{L^2(P_X)} = \frac{1}{2}\int_{\mathcal{X}}(F_\rho(x) - f(x))^2\, P_X(\mathrm{d}x)$, with residual $r_\rho := F_\rho - f$.

Definition 22. For a sample $(x_i, y_i)_{i=1}^N$ the *empirical risk* of $\rho \in \mathcal{P}(\Theta)$ is $L_N(\rho) := \frac{1}{2N}\sum_{i=1}^N (F_\rho(x_i) - y_i)^2$.

Definition 23. Let $\lambda, \beta > 0$ and $\nu_0 \in \mathcal{P}(Z)$ (in the paper $\nu_0(\mathrm{d}z) = Z_0^{-1}e^{-V(z)/\beta}\,\mathrm{d}z$). With $U(\theta) = \frac{\lambda}{2}a^2 + V(z)$ the *Gibbs reference measure* is $\mu_U(\mathrm{d}\theta) := Z_U^{-1}e^{-U(\theta)/\beta}\,\mathrm{d}\theta = \mathcal{N}(0, \beta/\lambda)(\mathrm{d}a) \otimes \nu_0(\mathrm{d}z)$.

Definition 24. The *domain* of the free energy is $\mathcal{D} := \{\rho \in \mathcal{P}(\Theta) : \mathrm{KL}(\rho\|\mu_U) < \infty\}$. Elements of $\mathcal{D}$ are not required to satisfy $\int V\,\mathrm{d}\rho < \infty$.

Definition 25. The *free energy* on $\mathcal{D}$ is $\mathcal{F}(\rho) := L(\rho) + \beta\ \mathrm{KL}(\rho\|\mu_U)$ (the paper's definition contains the additional constant $-\beta \log Z_U$, which is dropped here). By Lemma 183, whenever $\int U\,\mathrm{d}\rho < \infty$ one has $\mathcal{F}(\rho) = L(\rho) + \int U\,\mathrm{d}\rho + \beta\,\mathrm{Ent}(\rho) - \beta \log Z_U$.

Definition 26. The *empirical free energy* $\mathcal{F}_N$ is obtained from $\mathcal{F}$ by replacing L with L_N: $\mathcal{F}_N(\rho) := L_N(\rho) + \beta\ \mathrm{KL}(\rho\|\mu_U)$.

Definition 27. For a reference measure μ on Θ, a potential $W\colon \Theta \to \mathbb{R}$ and $\beta > 0$ with $Z_W := \int e^{-W/\beta}\,\mathrm{d}\mu < \infty$, the *Gibbs measure* is $\hat{\mu}_W(\mathrm{d}\theta) := Z_W^{-1}e^{-W(\theta)/\beta}\,\mu(\mathrm{d}\theta)$. (The paper takes μ to be Lebesgue measure on $\Theta = \mathbb{R} \times Z$.)

Theorem 28. *For $\nu_0 \in \mathcal{P}(Z)$ the Gibbs reference measure $\mu_U = \mathcal{N}(0, \beta/\lambda) \otimes \nu_0$ is a probability measure on Θ.*

Proof. $\square$

Definition 29. For measurable $s\colon Z \to \mathbb{R}$, $\lambda, \beta > 0$, the *conditional amplitude kernel* is $\kappa_s(z, \mathrm{d}a) := \mathcal{N}(-s(z)/\lambda, \beta/\lambda)(\mathrm{d}a)$.

2.5 Finite width and finite sample

Definition 30. For particles $\theta = (\theta_\ell)_{\ell=1}^M$, $\theta_\ell = (a_\ell, z_\ell)$, the *finite-width network* (mean-field scaling) is $F_\theta(x) := \frac{1}{M}\sum_{\ell=1}^M a_\ell \varphi_{z_\ell}(x)$.

Definition 31. For a sample $(x_i, y_i)_{i=1}^N$, $\lambda \geq 0$ and a potential $V\colon Z \to \mathbb{R}$, the *finite-width objective* is $J_N(\theta) := L_N(\theta) + \frac{1}{M}\sum_{\ell=1}^M U(\theta_\ell)$, where $L_N(\theta) = \frac{1}{2N}\sum_{i=1}^N (F_\theta(x_i) - y_i)^2$ and $U(a, z) = \frac{\lambda}{2}a^2 + V(z)$.

Definition 32. The *empirical inner product* on $\mathbb{R}^N$ is $\langle u, v\rangle_N := \frac{1}{N}\sum_{i=1}^N u_i v_i$, with norm $\|u\|_N := \langle u, u\rangle_N^{1/2}$.

Definition 33. For $r \in \mathbb{R}^N$ the *empirical analysis map* is $(S_N^* r)(z) := \frac{1}{N}\sum_{i=1}^N \varphi_z(x_i) r_i = \langle \varphi_z, r\rangle_N$ (independent of ν).

Definition 34. For a finite measure ν on Z (in the paper $\nu \in \mathcal{P}(Z)$) the *empirical kernel operator* $K_{N,\nu}$ on $(\mathbb{R}^N, \langle\cdot,\cdot\rangle_N)$ is $(K_{N,\nu}h)_i := \int_Z \varphi_z(x_i)\langle \varphi_z, h\rangle_N\,\nu(\mathrm{d}z)$; for $\nu_M = \frac{1}{M}\sum_\ell \delta_{z_\ell}$, $(K_{N,\nu_M}h)_i = \frac{1}{M}\sum_\ell \varphi_{z_\ell}(x_i)\langle \varphi_{z_\ell}, h\rangle_N$.

Definition 35. For fixed hidden parameters $z_1, \dots, z_M \in Z$ and the sample $(x_i)_{i=1}^N$, the *fixed-feature synthesis operator* is $S_N: (\mathbb{R}^M, \langle \cdot, \cdot \rangle_M) \to (\mathbb{R}^N, \langle \cdot, \cdot \rangle_N)$, $(S_N \gamma)_i := \frac{1}{M} \sum_{\ell=1}^M \gamma_\ell \varphi_{z_\ell}(x_i)$, with adjoint $(S_N^* r)_\ell = \frac{1}{N} \sum_i \varphi_{z_\ell}(x_i) r_i = (S_N^* r)(z_\ell)$.

Definition 36. For particles $\theta \in \Theta^M$ and the sample $(x_i, y_i)_{i=1}^N$ the *residual* is $r_i := F_\theta(x_i) - y_i$, $r = (r_i)_i \in \mathbb{R}^N$.

Theorem 37. *For $\theta \in \Theta^M$ with hidden empirical law $\nu_M = \frac{1}{M} \sum_\ell \delta_{z_\ell}$ and $h \in \mathbb{R}^N$, $(K_{N,\nu_M} h)_i = \frac{1}{M} \sum_{\ell=1}^M \varphi_{z_\ell}(x_i) \langle \varphi_{z_\ell}, h \rangle_N$.*

Proof. □

Theorem 38. *For $\lambda > 0$ the operator $K_{N,\nu_M} + \lambda$ is injective (hence bijective) on $\mathbb{R}^N$, since K_{N,ν_M} is symmetric positive semidefinite for $\langle \cdot, \cdot \rangle_N$.*

Proof. □

Definition 39. For $\lambda > 0$, $K_{N,\nu_M} + \lambda: \mathbb{R}^N \to \mathbb{R}^N$ is a linear bijection; $(K_{N,\nu_M} + \lambda)^{-1}$ denotes its inverse.

Definition 40. The vector field of the gradient flow $\dot{\theta} = -M \nabla J_N(\theta)$ is, with $r = (F_\theta(x_i) - y_i)_i$, $\theta \mapsto (-[(S_N^* r)(z_\ell) + \lambda a_\ell], -[a_\ell \nabla_z (S_N^* r)(z_\ell) + \nabla V(z_\ell)])_{\ell=1}^M$.

Definition 41. A curve $\theta: [0, \infty) \to \Theta^M$ is a solution of the gradient flow **(??)**, $\dot{\theta}_t = -M \nabla J_N(\theta_t)$, if it is differentiable on $[0, \infty)$ with $\dot{a}_\ell = -[(S_N^* r_t)(z_\ell) + \lambda a_\ell]$ and $\dot{z}_\ell = -[a_\ell \nabla_z (S_N^* r_t)(z_\ell) + \nabla V(z_\ell)]$.

Definition 42. For $\rho \in \mathcal{P}(\Theta)$ with $\int |a| \, d\rho < \infty$ (in particular for $\rho \in \mathcal{D}$), the realization F_ρ is a bounded measurable function, hence an element of $L^2(P_X)$.

Definition 43. For $f \in L^2(P_X)$ and ρ with $\int |a| \, d\rho < \infty$ the *residual* is $r_\rho := F_\rho - f \in L^2(P_X)$.

Definition 44. $\mathcal{H}_M := (\mathbb{R}^M, \langle \cdot, \cdot \rangle_M)$ with $\langle \gamma, \gamma' \rangle_M := \frac{1}{M} \sum_{\ell=1}^M \gamma_\ell \gamma'_\ell$.

Definition 45. The adjoint of S_N is $S_N^*: (\mathbb{R}^N, \langle \cdot, \cdot \rangle_N) \to (\mathbb{R}^M, \langle \cdot, \cdot \rangle_M)$, $(S_N^* r)_\ell = \frac{1}{N} \sum_i \varphi_{z_\ell}(x_i) r_i = (S_N^* r)(z_\ell)$.

Theorem 46. *For all $\gamma \in \mathbb{R}^M$ and $r \in \mathbb{R}^N$, $\langle S_N \gamma, r \rangle_N = \langle \gamma, S_N^* r \rangle_M$.*

Proof. □

Definition 47. S_N has finite rank, so $S_N^\dagger$ is the Moore–Penrose inverse and $S_N^\dagger y = (S_N^* S_N)^\dagger S_N^* y \in (\ker S_N)^\perp$; in Lean $S_N^\dagger y$ is the unique $u \in (\ker S_N)^\perp$ with $S_N^* S_N u = S_N^* y$.

Definition 48. The fixed-feature gradient flow with weight decay $\lambda \geq 0$ and no noise is $\dot{\gamma}_t = -S_N^*(S_N \gamma_t - y) - \lambda \gamma_t$, $\gamma_{t=0} = \gamma_0$ **(??)**.

Definition 49. For $\lambda > 0$, $\gamma_\lambda := (B + \lambda)^{-1} S_N^* y$, where $B = S_N^* S_N$.

2.6 Samples, empirical features and the risk with labels

Definition 50. For a sample $x_1, \dots, x_N \in \mathcal{X}$ the *empirical feature* is the feature map $(z, i) \mapsto \varphi_z(x_i)$ on the index space $\{1, \dots, N\}$, so that $\Phi_z = (\varphi_z(x_i))_{i=1}^N \in \mathbb{R}^N$.

Definition 51. The uniform probability measure on $\{1, \dots, N\}$; the empirical inner product is $\langle u, v\rangle_N = \frac{1}{N}\sum_{i=1}^N u_i v_i$ and $\|u\|_N^2 = \langle u, u\rangle_N$.

Definition 52. For the law Q of (X, Y) on $\mathcal{X} \times \mathbb{R}$ the *risk with labels* is $\tilde{L}(\rho) := \frac{1}{2}\,\mathbb{E}\big(F_\rho(X) - Y\big)^2$.

Definition 53. For a sample $\omega = (x_i, y_i)_{i=1}^N \in (\mathcal{X}\times\mathbb{R})^N$ the empirical risk is $L_N(\rho) = \frac{1}{2N}\sum_i (F_\rho(x_i) - y_i)^2$, a random functional of the sample.

Definition 54. $\mathcal{F}_N(\rho) := L_N(\rho) + \beta\ \mathrm{KL}(\rho\|\mu_U)$ (the constant $-\beta\log Z_U$ of the paper is dropped).

Definition 55. The sample $(x_i, y_i)_{i=1}^N$ consists of i.i.d. copies of $(X, Y) \sim Q$; its law is the product measure $Q^{\otimes N}$ on $(\mathcal{X} \times \mathbb{R})^N$.

Definition 56. $f : \mathcal{X} \to \mathbb{R}$ is a *regression function* of Q, $f(x) = \mathbb{E}[Y \mid X = x]$, if f is measurable and $\mathbb{E}\big[(Y - f(X))\, g(X)\big] = 0$ for every bounded measurable g (the tower property).

Theorem 57. *The uniform probability on $\{1, \dots, N\}$ gives mass $1/N$ to every point.*

Proof. □

Theorem 58. *For $g\colon \{1, \dots, N\} \to \mathbb{R}$, $\int g\ \mathrm{d}(\textit{uniform}) = \frac{1}{N}\sum_{i=1}^N g_i$.*

Proof. □

Definition 59. The identification of $\mathbb{R}^N$ with L^2(uniform): a vector $u \in \mathbb{R}^N$ is the function $i \mapsto u_i$.

Theorem 60. *For $u, v \in \mathbb{R}^N$, $\langle u, v\rangle_{L^2(\textit{uniform})} = \frac{1}{N}\sum_i u_i v_i = \langle u, v\rangle_N$.*

Proof. □

Theorem 61. *For $u \in \mathbb{R}^N$, $\|u\|_{L^2(\textit{uniform})} = \|u\|_N$.*

Proof. □

Theorem 62. *If $|y_i| \le Y_{\max}$ for all i then $\|y\|_N \le Y_{\max}$.*

Proof. □

Definition 63. For $\rho \in \mathcal{P}(\Theta)$ and the sample $(x_i, y_i)_{i=1}^N$ the *empirical residual* is $r_{N,\rho} := F_\rho|_x - y = (F_\rho(x_i) - y_i)_i \in \mathbb{R}^N$.

Definition 64. $\tilde{\mathcal{F}}(\rho) := \tilde{L}(\rho) + \beta\ \mathrm{KL}(\rho\|\mu_U)$ (the paper's constant $-\beta\log Z_U$ is dropped); $\tilde{\mathcal{F}} = \mathcal{F} + \text{const}$ has the same minimizer ρ^* as $\mathcal{F}$.

2.7 Rademacher complexity

Definition 65. For a class $\mathcal{G} = \{g_i : i \in \iota\}$ of functions on Ξ and points $\xi_1, \dots, \xi_N \in \Xi$, with i.i.d. Rademacher signs $\sigma_k \in \{\pm 1\}$, the *empirical Rademacher complexity* is $\widehat{\mathfrak{R}}_N(\mathcal{G}) := \mathbb{E}_\sigma \sup_i \left| \frac{1}{N} \sum_{k=1}^N \sigma_k g_i(\xi_k) \right|$ (absolute convention, FoML's 'empiricalRademacherComplexity'); the one-sided version $\mathbb{E}_\sigma \sup_i \frac{1}{N} \sum_k \sigma_k g_i(\xi_k)$ of the paper is 'empiricalRademacherOneSided'. For $\mathcal{G} = -\mathcal{G}$ the two agree.

Definition 66. The one-sided empirical Rademacher complexity $\mathbb{E}_\sigma \sup_i \frac{1}{N} \sum_{k=1}^N \sigma_k g_i(\xi_k)$ (without absolute value), FoML's 'empiricalRademacherComplexity_without_abs'.

Definition 67. The *Rademacher complexity* is the expectation of the empirical one over N i.i.d. points, $\mathfrak{R}_N(\mathcal{G}) := \mathbb{E}\, \widehat{\mathfrak{R}}_N(\mathcal{G})$ (FoML's 'rademacherComplexity').

Definition 68. The *feature class* of a feature map φ is $\Phi := \{\varphi_z : z \in Z\}$, a class of functions on $\mathcal{X}$ indexed by Z.

Definition 69. For points $x_1, \dots, x_N \in \mathcal{X}$, $\widehat{\mathfrak{R}}_N(\Phi) := \mathbb{E}_\sigma \sup_{z \in Z} \left| \frac{1}{N} \sum_{k=1}^N \sigma_k \varphi_z(x_k) \right|$. For $\Phi = -\Phi$ (as for tanh) this is the one-sided complexity of the paper.

Definition 70. For the law Q of (X, Y), $\mathfrak{R}_N(\Phi) := \mathbb{E}\, \widehat{\mathfrak{R}}_N(\Phi)$, the expectation over the inputs $x_1, \dots, x_N$ of an i.i.d. sample of size N from Q.

Definition 71. The *product class* of the feature map is $\Psi := \{\varphi_z \varphi_{z'} : z, z' \in Z\}$, indexed by $Z \times Z$; its members are bounded by one.

Definition 72. $\widehat{\mathfrak{R}}_N(\Psi) := \mathbb{E}_\sigma \sup_{z, z' \in Z} \left| \frac{1}{N} \sum_{k=1}^N \sigma_k \varphi_z(x_k) \varphi_{z'}(x_k) \right|$, and $\mathfrak{R}_N(\Psi) := \mathbb{E}\, \widehat{\mathfrak{R}}_N(\Psi)$ over the inputs of an i.i.d. sample from Q.

Definition 73. $\mathfrak{R}_N(\Psi) := \mathbb{E}\, \widehat{\mathfrak{R}}_N(\Psi)$, the expectation over the inputs $x_1, \dots, x_N$ of an i.i.d. sample of size N from Q.

Definition 74. For $B \geq 0$, $\mathcal{P}_B := \{\rho \in \mathcal{P}(\Theta) : \int |a|\, \rho(\mathrm{d}a\, \mathrm{d}z) \leq B\}$ (the integrability of a is part of the definition).

Definition 75. For a sample $\omega = (x_i, y_i)_{i=1}^N$, $G_+(\omega) := \sup_{\rho \in \mathcal{P}_B} (\tilde{L}(\rho) - L_N(\rho))$.

Definition 76. $G_-(\omega) := \sup_{\rho \in \mathcal{P}_B} (L_N(\rho) - \tilde{L}(\rho))$.

Definition 77. $G(\omega) := \sup_{\rho \in \mathcal{P}_B} |L_N(\rho) - \tilde{L}(\rho)| = \max(G_+, G_-)$.

Definition 78. For $e \colon \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ (in the paper $e(x, y) = y - f(x)$), $G_N(\omega) := \sup_{z \in Z} \left| \frac{1}{N} \sum_{i=1}^N \varphi_z(x_i) e(x_i, y_i) - \mathbb{E}[\varphi_z(X) e(X, Y)] \right|$.

Definition 79. $H_N(\omega) := \sup_{z, z' \in Z} \left| \frac{1}{N} \sum_{i=1}^N \varphi_z(x_i) \varphi_{z'}(x_i) - \mathbb{E}[\varphi_z(X) \varphi_{z'}(X)] \right|$.

Definition 80. For a countable $Z_0 \subseteq Z$, $\mathcal{P}_B^0 := \left\{ \frac{1}{M} \sum_{j=1}^M \delta_{(\epsilon_j q, z_j)} : M \geq 1,\ q \in \mathbb{Q},\ \epsilon_j \in \{\pm 1\},\ z_j \in Z_0 \right\} \cap \mathcal{P}_B$, a countable subclass of $\mathcal{P}_B$.

Definition 81. For finite measures ν_1, ν_2 and functions m_1, m_2, with $\mu := \nu_1 + \nu_2$ and $w_i := \mathrm{d}\nu_i / \mathrm{d}\mu$, the *mixed density* is $\psi := w_1 m_1 - w_2 m_2$, so that $m_1 \nu_1 - m_2 \nu_2 = \psi\, \mu$.

Definition 82. $B := \frac{Y_{\max}}{\lambda} + \sqrt{\frac{2\beta}{\pi\lambda}}$, the bound on $\int |a|\, \mathrm{d}\rho$ for both minimizers (Lemma **??**(c)).

Definition 83. $\Delta_N(\delta) := 4B(B+Y_{\max})\,\mathfrak{R}_N(\Phi) + (B+Y_{\max})^2\sqrt{\frac{\log(2/\delta)}{2N}}$.

Definition 84. $D(\delta) := 4C_{\mathrm{P}}B(B+Y_{\max})\sqrt{m+1} + (B+Y_{\max})^2\sqrt{\frac{1}{2}\log(2/\delta)}$.

Definition 85. $N_0(\varepsilon,\delta) := \lceil (D(\delta)/(\beta\varepsilon))^2 \rceil$.

Definition 86. $\Delta'_N := 2(B+Y_{\max})\,\mathfrak{R}_N(\Phi) + (B+Y_{\max})\sqrt{\frac{2\log(4/\delta)}{N}}$.

Definition 87. $E_\delta := \{\omega : |y_i| \le Y_{\max}\ \forall i,\ G(\omega) \le \frac{1}{2}\Delta_N(\delta)\}$, the event of Lemma 302 (with the full-measure condition on the labels).

Definition 88. The setting of Section **??**: $Q \in \mathcal{P}(\mathcal{X}\times\mathbb{R})$ with $|Y| \le Y_{\max}$ a.s., $f = \mathbb{E}[Y \mid X]$ a regression function with $|f| \le Y_{\max}$, $\lambda, \beta > 0$, and $\rho^* \in \mathcal{D}$ a minimizer of $\mathcal{F}$ on $\mathcal{D}$ (with target $f \in L^2(P_X)$).

Definition 89. For every sample $\omega = (x_i, y_i)_{i=1}^N$, $\rho^*_N(\omega) \in \mathcal{D}$ is a minimizer of $\mathcal{F}_N$ on $\mathcal{D}$ (Lemma **??** for L_N).

Definition 90. $e(x,y) := y - F_{\rho^*}(x)$ (the paper's e has the opposite sign; only $|e|$ enters).

Definition 91. $E'_\delta := E_{\delta/2} \cap \{G_N \le \Delta'_N\}$, where G_N is the deviation of Lemma 305 for $e(x,y) = y - F_{\rho^*}(x)$ and Δ'_N is (86).

Definition 92. For a sample $\omega = (x_i, y_i)_i$ and $\rho \in \mathcal{P}(\Theta)$, $g_{N,\rho} := S^*_N(F_\rho|_x - y)$, i.e. $g_{N,\rho}(z) = \frac{1}{N}\sum_i \varphi_z(x_i)(F_\rho(x_i) - y_i)$; for $\rho = \rho^*_N$ this is g_N of Lemma **??**.

Definition 93. For a sample $\omega = (x_i, y_i)_i$, $\nu \in \mathcal{P}(Z)$ and $\lambda > 0$, $R^N_{\lambda,\nu} y := S^*_N(K_{N,\nu} + \lambda)^{-1} y$, the regularized ridgelet transform of the empirical feature.

Definition 94. $D'(\delta) := (B+Y_{\max})[2C_{\mathrm{P}}\sqrt{m+1} + \sqrt{2\log(4/\delta)}]$, so that $\Delta'_N \le D'(\delta)/\sqrt{N}$.

Definition 95. $e(x,y) := F_{\rho^*}(x) - y$, a fixed (sample-independent) function on $\mathcal{X}\times\mathbb{R}$ with $\sup|e| \le E_* := M_* + Y_{\max}$.

Definition 96. $h := F_{\rho^*_N} - F_{\rho^*}$, with $\|h\|_N := \|h|_x\|_N$ and $\|h\| := \|h\|_{L^2(P_X)}$.

Definition 97. $\ell_F(x,y) := \frac{1}{2}(F(x) - y)^2$, so that $\tilde{L}(\rho) = P\ell_{F_\rho}$ and $L_N(\rho) = P_N \ell_{F_\rho}$.

Definition 98. For $g\colon \mathcal{X}\times\mathbb{R} \to \mathbb{R}$, $P_N g := \frac{1}{N}\sum_{i=1}^N g(x_i, y_i)$ and $Pg := \mathbb{E}g(X,Y)$.

Definition 99. $(P_N - P)g := P_N g - Pg$.

Definition 100. $\mathcal{K}_N := \mathrm{KL}(\rho^* \| \rho^*_N) + \mathrm{KL}(\rho^*_N \| \rho^*)$ (the Jeffreys divergence).

Definition 101. $D_\varsigma := |\Pi\varsigma|(Z)$, the total variation of the signed measure $\Pi\varsigma = \Pi\rho^*_N - \Pi\rho^*$ on Z.

Definition 102. $m^* := -g^*/\lambda$ with $g^* = S^* r^*$, the everywhere-defined continuous version of m_{ρ^*} (Theorem **??**(b)).

Definition 103. $m_N := -g_N/\lambda$ with $g_N = S^*_N r^*_N$, the everywhere-defined version of $m_{\rho^*_N}$ (Lemma **??**(a)).

Definition 104. With $\mu := \nu^*_N + \nu^*$, $w_N := \mathrm{d}\nu^*_N/\mathrm{d}\mu$, $w^* := \mathrm{d}\nu^*/\mathrm{d}\mu$, the density of $\Pi\varsigma = m_N\nu^*_N - m^*\nu^*$ with respect to μ is $\psi := w_N m_N - w^* m^*$, so that $D_\varsigma = \int |\psi|\,\mathrm{d}\mu$.

Definition 105. $\eta'_N := 2E_*\mathfrak{R}_N(\Phi) + E_*\sqrt{2\log(4/\delta)/N}$, the bound of Lemma 305 with $\delta_1 = \delta/2$; under Lemma 314, $\eta'_N \le E_* D'(\delta)/\sqrt{N}$ with $D'(\delta) = 2C_{\mathrm{P}}\sqrt{m+1} + \sqrt{2\log(4/\delta)}$.

Definition 106. $\eta''_N := 2\mathfrak{R}_N(\Psi) + \sqrt{2\log(4/\delta)/N}$, the bound of Lemma 303 with $\delta_2 = \delta/2$; under Lemma 315, $\eta''_N \le D''(\delta)/\sqrt{N}$ with $D''(\delta) = 2C'_{\mathrm{P}}\sqrt{m+1} + \sqrt{2\log(4/\delta)}$.

Definition 107. $c_1 := \max\{1, 1/\lambda, M_*/\sqrt{\beta}\}$ with $M_* = Y_{\max}/\lambda$.

Definition 108. For a bound $\sup_z |m^*(z)| \le M_1$ on the population coefficient, $c_1(M_1) := \max\{1, 1/\lambda, M_1/\sqrt{\beta}\}$; the constant c_1 of Theorem efthm:m3-double-prime is $c_1(M_*)$ with $M_* = Y_{\max}/\lambda$.

Definition 109. $\Omega_\delta := \{G_N \le \eta'_N\} \cap \{H_N \le \eta''_N\}$; $\mathbb{P}(\Omega_\delta) \ge 1 - \delta$ by Lemmas 305 and 303.

Theorem 110. *The pseudo-dimension of the class $\Phi = \{\varphi_z : z \in Z\}$ of* tanh *ridge functions $x \mapsto$* $\tanh(w^\top x - b)$*, $(w, b) \in \mathbb{R}^m \times \mathbb{R}$, is at most $m + 1$ (Lemma 314(a)); this is Lemma 811 for $E = \mathbb{R}^m$.*

Proof. □

Definition 111. For $A > 0$ and $c \in \mathbb{N}$, $\mathrm{HC}(m, A, c)$ is the statement: for every N, every $x_1, \dots, x_N \in \mathbb{R}^m$ and every $0 < \varepsilon \le 1/2$, $\mathcal{N}(\varepsilon, \Phi, L_2(P_N)) \le (A/\varepsilon)^{c(m+1)}$. Haussler's bound gives $\mathrm{HC}(m, A, 2)$ with an absolute A.

2.8 Synthesis of a signed measure

Definition 112. For a finite signed Borel measure γ on Z the synthesis is $(S\gamma)(x) := \int_Z \varphi_z(x)\,\gamma(\mathrm{d}z)$; if $\gamma = h\nu$ with $h \in L^2(\nu)$ then $S\gamma = S_\nu h$, and $F_\rho = S\,\Pi\rho$ when $\int |a|\,\mathrm{d}\rho < \infty$.

Definition 113. For a finite signed measure γ on Z, $L(\gamma) := \frac{1}{2}\|S\gamma - f\|^2_{L^2(P_X)}$.

2.9 Dynamics: log-Sobolev inequality, proximal Gibbs measure, mean-field Langevin flow

Definition 114. For $\rho \in \mathcal{D}$ let $r_\rho = F_\rho - f$, $s_\rho := S^* r_\rho$ and $W_\rho(\theta) := a\,s_\rho(z) + U(\theta)$. The *proximal Gibbs measure* is $\hat{\mu}_\rho(\mathrm{d}\theta) := Z_\rho^{-1} e^{-W_\rho(\theta)/\beta}\,\mathrm{d}\theta = Z_\rho^{-1} e^{-a s_\rho(z)/\beta}\mu_U(\mathrm{d}\theta)$, where $Z_\rho = \int e^{-W_\rho/\beta}\,\mathrm{d}\theta < \infty$ since $|s_\rho| \le c_\rho$.

Definition 115. A probability measure μ on $\mathbb{R}^d$ satisfies $\mathrm{LSI}(\alpha)$ with $\alpha > 0$ if for all $g \in C_c^\infty(\mathbb{R}^d)$

$$\mathrm{Ent}_\mu(g^2) := \int g^2 \log g^2\,\mathrm{d}\mu - \Big(\int g^2\,\mathrm{d}\mu\Big)\log\Big(\int g^2\,\mathrm{d}\mu\Big) \le \frac{2}{\alpha}\int |\nabla g|^2\,\mathrm{d}\mu.$$

(Here the test functions are C_c^1; by standard approximation the two conventions agree.)

Definition 116. ρ has a *smooth density* with respect to μ if $\rho = p\,\mu$ for a positive C^1 function p with $\nabla \log p \in L^2(\rho)$.

Definition 117. For $\rho \ll \mu$ with a smooth density $p = \frac{\mathrm{d}\rho}{\mathrm{d}\mu}$ the *relative Fisher information* is $I(\rho|\mu) := \int \left|\nabla \log \frac{\mathrm{d}\rho}{\mathrm{d}\mu}\right|^2 \mathrm{d}\rho = \int |\nabla \log p|^2\,\mathrm{d}\rho$.

Definition 118. A probability measure μ satisfies the *KL form of* LSI(α) if for every probability measure $\rho \ll \mu$ with a smooth density and $\mathrm{KL}(\rho\|\mu) < \infty$,

$$\mathrm{KL}(\rho\|\mu) \leq \frac{1}{2\alpha} I(\rho|\mu).$$

This is (L1) of the paper, obtained from LSI(α) by substituting $g = \sqrt{\mathrm{d}\rho/\,\mathrm{d}\mu}$.

Definition 119. For $s\colon Z \to \mathbb{R}$ and $\lambda > 0$ the *shear* is $T_s(a, z) := (a + s(z)/\lambda,\ z)$, a bijection of Θ with inverse $T_s^{-1}(a', z) = (a' - s(z)/\lambda,\ z)$.

Definition 120. The parameter space $\Theta = \mathbb{R} \times Z$ is given the Euclidean structure $|\theta|^2 = a^2 + |z|^2$; a measure on Θ is regarded as a measure on this Euclidean space (formally, its image under the identity map to 'WithLp 2 (ℝ × Z)').

Definition 121. For measures ρ, μ on Θ, $I(\rho|\mu)$ is the relative Fisher information with respect to the Euclidean gradient $\nabla_\theta = (\partial_a, \nabla_z)$.

Definition 122. A probability measure μ on $\Theta = \mathbb{R} \times Z$ satisfies LSI(α) if it does so as a measure on the Euclidean space Θ.

Definition 123. A probability measure μ on Θ satisfies the KL form of LSI(α) if $\mathrm{KL}(\rho\|\mu) \leq \frac{1}{2\alpha} I(\rho|\mu)$ for every probability measure ρ on Θ with a smooth density with respect to μ and $\mathrm{KL}(\rho\|\mu) < \infty$.

Definition 124. *(A8), Gaussian confinement.* $V(z) = \frac{\lambda_z}{2}|z|^2$ with $\lambda_z > 0$, so that $\nu_0(\,\mathrm{d}z) = Z_0^{-1} e^{-V(z)/\beta}\,\mathrm{d}z = \mathcal{N}(0, (\beta/\lambda_z)I)(\,\mathrm{d}z)$.

Definition 125. For $c \geq 0$ let $\alpha(c) := \dfrac{\min\{\lambda,\ \lambda_z e^{-c^2/(2\lambda\beta)}\}}{\beta\,(1 + \sqrt{R_X^2 + 1}\,c/\lambda)^2}$, the right-hand side of the LSI constant of $\hat{\mu}_\rho$ in Lemma **??**(iv) with $c = c_\rho$; it is nonincreasing in c.

Definition 126. For $E \in \mathbb{R}$ let $c_E := \sqrt{2(E + \beta \log Z_U)}$ (here $c_E = \sqrt{2E}$, the constant of the free energy being dropped) and

$$\alpha_*(E) := \frac{\min\{\lambda,\ \lambda_z\, e^{-c_E^2/(2\lambda\beta)}\}}{\beta\,(1 + \sqrt{R_X^2 + 1}\,c_E/\lambda)^2}.$$

Definition 127. The feature map is *smooth with bounded derivatives* if $z \mapsto \varphi_z(x)$ is C^∞ for every x and $\sup_{x,z} \|\nabla_z^n \varphi_z(x)\| < \infty$ for every $n \geq 1$. Under (A1) and (A3) this holds with $\sup_{x,z} \|\nabla_z \varphi_z(x)\| \leq \sqrt{R_X^2 + 1}$ and $\sup_{x,z} \|\nabla_z^2 \varphi_z(x)\| \leq c_2(R_X^2 + 1)$.

Definition 128. *(Hypothesis W as the definition of a solution in law.)* A curve $(\rho_t)_{t\geq 0}$ of probability measures on Θ is an *MFLD flow* if

(W1) $\rho_t \in \mathcal{D}$ for all $t \geq 0$, $t \mapsto \rho_t$ is weakly continuous, $\sup_{t\leq T} \int |\theta|^2\,\mathrm{d}\rho_t < \infty$ for every T, and for a.e. $t \geq 0$ the measure ρ_t has a smooth density with respect to $\hat{\mu}_{\rho_t}$ with $\nabla \log \frac{\mathrm{d}\rho_t}{\mathrm{d}\hat{\mu}_{\rho_t}} \in L^2(\rho_t)$;

(W2) $u \mapsto I(\rho_u|\hat{\mu}_{\rho_u})$ is locally integrable on $[0, \infty)$ and for all $0 \leq s \leq t$

$$\mathcal{F}(\rho_t) = \mathcal{F}(\rho_s) - \beta^2 \int_s^t I(\rho_u|\hat{\mu}_{\rho_u})\,\mathrm{d}u.$$

(W2) is the integrated form of $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(\rho_t) = -\beta^2 I(\rho_t|\hat{\mu}_{\rho_t})$ for a locally absolutely continuous $t \mapsto \mathcal{F}(\rho_t)$.

2.10 The finite particle system: Gibbs measure and free energy

Definition 129. For $\theta = (\theta_\ell)_{\ell=1}^M \in \Theta^M$ the *particle risk* is $L(\rho_\theta) = \frac{1}{2}\|F_{\rho_\theta} - f\|^2_{L^2(P_X)}$, where $\rho_\theta = \frac{1}{M}\sum_\ell \delta_{\theta_\ell}$ and $F_{\rho_\theta}(x) = \frac{1}{M}\sum_\ell a_\ell \varphi_{z_\ell}(x)$.

Definition 130. The *M-particle Gibbs measure* is $\pi_M(\mathrm{d}\theta) := Z_M^{-1} e^{-ML(\rho_\theta)/\beta}\, \mu_U^{\otimes M}(\mathrm{d}\theta)$, $Z_M := \int e^{-ML(\rho_\theta)/\beta}\, \mathrm{d}\mu_U^{\otimes M}$. Since $\mu_U^{\otimes M} \propto e^{-\sum_\ell U(\theta_\ell)/\beta}\,\mathrm{d}\theta$, this is $\pi_M \propto e^{-MJ(\theta)/\beta}\,\mathrm{d}\theta$ with $J(\theta) = L(\rho_\theta) + \frac{1}{M}\sum_\ell U(\theta_\ell)$, the invariant measure of the M-particle noisy gradient descent.

Definition 131. For $\mu \in \mathcal{P}(\Theta^M)$ the *M-particle free energy* is $\mathcal{F}^M(\mu) := M\int_{\Theta^M} L(\rho_\theta)\,\mu(\mathrm{d}\theta) + \beta\,\mathrm{KL}(\mu\|\mu_U^{\otimes M})$ (the paper's $\mathcal{F}^M$ up to the dropped constant $-\beta\log Z_{\pi_M}$).

Definition 132. For $\rho \in \mathcal{P}(\Theta)$, $\rho^{\otimes M} \in \mathcal{P}(\Theta^M)$ is the law of M i.i.d. particles of law ρ; the comparison measure of static chaos is $\rho^{*\otimes M}$ for the minimizer ρ^* of $\mathcal{F}$.

Definition 133. **(E)** A family $(\mu_t)_{t\geq 0} \subset \mathcal{P}(\Theta^M)$ is *ergodic to* $\pi \in \mathcal{P}(\Theta^M)$ if $\|\mu_t - \pi\|_{\mathrm{TV}} \to 0$ as $t \to \infty$ and $\sup_{t\geq 0}\int \frac{1}{M}\sum_\ell a_\ell^2\,\mu_t(\mathrm{d}\theta) < \infty$. For the laws of the M-particle noisy gradient descent and $\pi = \pi_M$ the first property is the ergodic theorem for non-degenerate Langevin diffusions with an invariant probability measure and the second is the moment bound of the particle system.

Definition 134. For a sample $\omega = (x_i, y_i)_{i=1}^N$ the *empirical M-particle Gibbs measure* is $\pi_M(\mathrm{d}\theta) \propto e^{-ML_N(\rho_\theta)/\beta}\mu_U^{\otimes M}(\mathrm{d}\theta)$, the invariant measure of the M-particle empirical noisy gradient descent ($J = J_N$); it is the particle Gibbs measure of the empirical feature for the uniform measure on $\{1,\dots,N\}$ and the target y (Lemma 271).

Definition 135. $C_g(Y_{\max},\lambda,\beta) := \frac{1}{2\beta}\left(\frac{{Y_{\max}}^2}{\lambda^2} + \frac{\beta}{\lambda}\right) + \log 2 + 4\exp\left(2C\left(\frac{Y_{\max}}{\lambda} + \sqrt{\frac{2\beta}{\pi\lambda}}\right)\right)\exp\left(\frac{\beta}{2\lambda}(2C + \frac{Y_{\max}}{\beta})^2\right)$, a bound on the constant C_g^N of Theorem 550 that depends on the sample only through $Y_{\max}$ ($A_N \leq {Y_{\max}}^2/\lambda^2 + \beta/\lambda$, and (**??**) with $\|f\| \to Y_{\max}$).

2.11 Convergence rates: the weighted Tikhonov problem

Definition 136. On $L^2(\nu_0)$ let $T_0 := S_{\nu_0}^* S_{\nu_0}$. It is positive semidefinite with $\langle T_0 u, u\rangle = \|S_{\nu_0}u\|^2$ and $\|T_0\| \leq 1$.

Theorem 137. *$T_0 = S_{\nu_0}^* S_{\nu_0}$ is self-adjoint and positive semidefinite with $\langle T_0 u, u\rangle = \|S_{\nu_0}u\|^2$, and $\|T_0\| \leq 1$.*

Proof. □

Definition 138. For $\lambda > 0$ and $f \in L^2(P_X)$, $u_\lambda := (T_0 + \lambda)^{-1} S_{\nu_0}^* f = S_{\nu_0}^*(K_{\nu_0} + \lambda)^{-1} f = R_{\lambda,\nu_0} f$; the two expressions agree by the push-through identity (Lemma **??**(4)).

Definition 139. **(SC$_a$)** *Source condition of order* $a > 0$. There is $g_0 \in L^2(\nu_0)$ with $u^\dagger = T_0^a g_0$. Since $\mathrm{ran}\, T_0^a \subset \overline{\mathrm{ran}\, T_0} = (\ker S_{\nu_0})^\perp$, such a $u^\dagger$ automatically satisfies the minimum-norm condition, and (SC$_a$) contains (R) ($f = S_{\nu_0} u^\dagger$). For $a = 1/2$, $\mathrm{ran}\, T_0^{1/2} = \mathrm{ran}\, S_{\nu_0}^*$, so (SC$_{1/2}$) is equivalent to $u^\dagger = S^* g$ with $g \in L^2(P_X)$, i.e. to $f = K_{\nu_0} g$. (In Lean the condition is formalized for $a \in \{1/2, 1\}$ with a bound G on the norm of the source element: $f = K_{\nu_0} g$, $\|g\| \leq G$, resp. $f = S_{\nu_0} T_0 g_0$, $\|g_0\| \leq G$; the spectral power T_0^a of an operator on a real Hilbert space is not available in Mathlib.)

Definition 140. **($\mathbf{S}_\infty$)** *Uniform sup-norm bound.* For the family of (λ, β) under consideration (for instance $\lambda \le \lambda_0$, $\kappa \le \kappa_0$), $\sup_{(\lambda,\beta)} \sup_{z\in Z} |m^*_{\lambda,\beta}(z)| \le B_\infty < \infty$, where $m^*_{\lambda,\beta}$ is the conditional mean amplitude of the minimizer of $\mathcal{F}_{\lambda,\beta}$.

Definition 141. $v := w\,m^* \in L^2(\nu_0)$, where $w = \mathrm{d}\nu^*/\,\mathrm{d}\nu_0$ and $m^* = m_{\rho^*}$; v is bounded, and $\Pi\rho^* = m^*\nu^* = v\,\nu_0$.

Definition 142. $\Xi := \|(w-1)m^*\|_{L^2(\nu_0)}$, where $w = \mathrm{d}\nu^*/\,\mathrm{d}\nu_0$ and $m^* = m_{\rho^*}$; Ξ is the whole contribution of the learned base measure to the rate.

Definition 143. $\lambda_N := N^{-1/p}$ for $p > 0$; $p = 4(a+2)$ in Theorem **??**, $p = 2(a+3)$ in Corollary **??**(iii).

Definition 144. $C_{\mathrm{stat}}(m,\delta) := \sqrt{2[4C_{\mathrm{P}}\sqrt{m+1} + \sqrt{\tfrac{1}{2}\log\tfrac{4}{\delta}}]} + 2C_{\mathrm{P}}\sqrt{m+1} + \sqrt{2\log\tfrac{4}{\delta}}$.

Definition 145. $C_1 := \tfrac{1}{2}(B_\infty^2 e^{\kappa_0 B_\infty^2/2} + \|u^\dagger\|^2)e^{\kappa_0\|u^\dagger\|^2/4}\|u^\dagger\|$.

Definition 146. $C_1' := \tfrac{1}{2}(B_\infty^2 e^{\kappa_0 B_\infty^2/2} + \|u^\dagger\|^2)$.

Definition 147. With $\zeta_{0,0} := \frac{\kappa_0}{2}\|u^\dagger\|^2$, $C_2 := \|u^\dagger\|^3 e^{\zeta_{0,0}/2}\max\Big\{\frac{e^{\zeta_{0,0}}}{8}\exp\Big(\frac{e^{\zeta_{0,0}}\|u^\dagger\|^2}{8\beta_0}\Big), \frac{1}{2}\Big\}$.

Definition 148. **(R)** $f \in \operatorname{ran} S_{\nu_0}$, that is, there is $u_0 \in L^2(\nu_0)$ with $f = S_{\nu_0}u_0 = \int_Z u_0(z)\varphi_z\,\nu_0(\mathrm{d}z)$ in $L^2(P_X)$. Such a u_0 is called a *representer* of f (with respect to ν_0).

Definition 149. Under (R), the *canonical (minimum-norm) ridgelet transform* of f with respect to ν_0 is $u_0^\dagger := S_{\nu_0}^\dagger f := P_{(\ker S_{\nu_0})^\perp}u_0$, the orthogonal projection onto $(\ker S_{\nu_0})^\perp$ of any representer u_0 of f (Lemma 382: it does not depend on u_0). Even if $\operatorname{ran} S_{\nu_0}$ is not closed, $S_{\nu_0}^\dagger f$ is defined in this sense as long as $f \in \operatorname{ran} S_{\nu_0}$. (In Lean, $u_0^\dagger := 0$ when $f \notin \operatorname{ran} S_{\nu_0}$.)

Definition 150. For $\nu \in \mathcal{P}(Z)$, $u\colon Z \to \mathbb{R}$ measurable and $\sigma^2 := \beta/\lambda$ put $\rho_{\nu,u} := \nu(\mathrm{d}z) \otimes \mathcal{N}(u(z),\sigma^2)(\mathrm{d}a)$, a probability measure on Θ with hidden marginal ν and conditional mean amplitude u. The *competitor* of the paper is $\tilde\rho_u := \rho_{\nu_0,u}$ for $u \in L^2(\nu_0)$.

Definition 151. For ρ with $\int|a|\,\mathrm{d}\rho < \infty$, $\lambda > 0$ and a finite measure ν on Z, the bounded measurable function $-s_\rho/\lambda = -\lambda^{-1}S^*r_\rho$ is an element of $L^2(\nu)$; for $\nu = \nu^*$ it is m_{ρ^*} (Theorem 228).

Definition 152. $\tilde Z := \int_Z \exp(\frac{\kappa}{2}m^*(z)^2)\nu_0(\mathrm{d}z) = \int_Z \exp(\frac{s^*(z)^2}{2\lambda\beta})\nu_0(\mathrm{d}z) = Z_*/Z_0$ in the notation of Theorem 226.

Definition 153. For ρ with $\int|a|\,\mathrm{d}\rho < \infty$ and $\lambda > 0$, $m_\rho := -s_\rho/\lambda = -\lambda^{-1}S^*r_\rho\colon Z \to \mathbb{R}$, a bounded measurable function; for the minimizer ρ^* it is the conditional mean amplitude (Theorem 228).

Definition 154. $w := \frac{\mathrm{d}\nu^*}{\mathrm{d}\nu_0} = \frac{Z_0}{Z_*}\exp(\frac{\kappa}{2}m^{*2}) = \tilde Z^{-1}\exp(\frac{\kappa}{2}m^{*2})$ with $\kappa = \lambda/\beta$ (Theorem 401).

Definition 155. Assume (A1), (A3), (A5), $f \in L^2(P_X)$, (R) and the convention (**??**) (ν_0 fixed). Let $(\lambda_n,\beta_n)_{n\ge1}$ satisfy $\lambda_n \to 0$ and $\kappa_n := \lambda_n/\beta_n \to 0$, and let ρ_n^* be the minimizer of $\mathcal{F}$ for (λ_n,β_n); write $\nu_n^* := \nu_{\rho_n^*}$, $m_n^* := m_{\rho_n^*}$, $\gamma_n := \Pi\rho_n^* = m_n^*\nu_n^*$ and $\gamma_\infty := u_0^\dagger\nu_0$.

Definition 156. $r_N := \frac{1}{2\lambda}\Big[\sqrt{2D(\delta/2)/\sqrt{N}} + D'(\delta)/\sqrt{N}\Big] + \frac{Y_{\max}}{\lambda}\sqrt{\frac{D(\delta/2)}{2\beta\sqrt{N}}}$, the bound of Corollary 340 with $\Delta_N \le D(\delta/2)/\sqrt{N}$ and $\Delta_N' \le D'(\delta)/\sqrt{N}$ (Corollary 341); $r_N \to 0$ as $N \to \infty$.

Definition 157. $\mathcal{P}_2 := \{\rho \in \mathcal{P}(\Theta) : \int a^2 \, \mathrm{d}\rho < \infty\}$.

Definition 158. For $\gamma \neq 0$ with $c := \|\gamma\|_{\mathcal{M}}$ and Jordan decomposition $\gamma = \gamma^+ - \gamma^-$, the lift $\bar{\rho} := c^{-1}(\gamma^+(\mathrm{d}z) \otimes \delta_c(\mathrm{d}a) + \gamma^-(\mathrm{d}z) \otimes \delta_{-c}(\mathrm{d}a))$ is the measure $\frac{|\gamma|(\mathrm{d}z)}{c} \otimes \delta_{c\vartheta(z)}(\mathrm{d}a)$ of (159), with $\vartheta = \mathrm{d}\gamma/\mathrm{d}|\gamma| \in \{\pm 1\}$.

Definition 159. For $\gamma \neq 0$, $\bar{\rho}(\mathrm{d}a \, \mathrm{d}z) := \frac{|\gamma|(\mathrm{d}z)}{c} \otimes \delta_{c\,\vartheta(z)}(\mathrm{d}a)$ with $c = \|\gamma\|_{\mathcal{M}}$ and $\vartheta = \mathrm{d}\gamma/\mathrm{d}|\gamma| \in \{\pm 1\}$; for $\gamma = 0$, $\bar{\rho} := \delta_0 \otimes \nu$ with an arbitrary $\nu \in \mathcal{P}(Z)$.

Definition 160. The functional of the $\beta = 0$ problem (**??**) is $\mathcal{E}_0(\rho) := L(\rho) + \frac{\lambda}{2} \int a^2 \, \mathrm{d}\rho$ on $\mathcal{P}_2$.

Definition 161. The total-variation regularized functional on $\mathcal{M}(Z)$ is $\mathcal{E}_{TV}(\gamma) := L(\gamma) + \frac{\lambda}{2}\|\gamma\|_{\mathcal{M}}^2$.

Theorem 162. *S_{ν_0} is Hilbert–Schmidt with $\|S_{\nu_0}\|_{HS}^2 = \sum_i \|S_{\nu_0} e_i\|^2 \leq 1$ along any Hilbert basis (e_i) of $L^2(\nu_0)$: with $\varphi_x := \varphi_\cdot(x) \in L^2(\nu_0)$, $\|S_{\nu_0} u\|^2 = \int \langle u, \varphi_x\rangle^2 \, P_X(\mathrm{d}x)$, and Bessel's inequality gives $\sum_{i \in s} \langle e_i, \varphi_x\rangle^2 \leq \|\varphi_x\|^2 \leq 1$ for every finite s. Consequently $\sum_i \langle T_0 e_i, e_i\rangle = \sum_i \|S_{\nu_0} e_i\|^2 \leq 1$ and T_0 is compact.*

Proof. □

Theorem 163. *T_0 is a compact positive operator on the separable Hilbert space $L^2(\nu_0)$, hence there is a countable Hilbert basis (e_k) of $L^2(\nu_0)$ with $T_0 e_k = \mu_k e_k$, $\mu_k = \langle T_0 e_k, e_k\rangle = \|S_{\nu_0} e_k\|^2 \in [0, 1]$.*

Proof. □

Definition 164. For $a \geq 0$, $T_0^a := \sum_k \mu_k^a \langle e_k, \cdot\rangle e_k$ is the spectral power of T_0 (Definition 789 with respect to the eigenbasis of Lemma 163); it does not depend on the choice of the eigenbasis (Lemma 794).

Definition 165. **(SC$_a$)** *Source condition of order $a > 0$.* There is $g_0 \in L^2(\nu_0)$ with $u^\dagger = T_0^a g_0$, where T_0^a is the spectral power of Definition 164. Since ran $T_0^a \subset (\ker S_{\nu_0})^\perp$, such a $u^\dagger$ automatically satisfies the minimum-norm condition, and (SC$_a$) contains (R) ($f = S_{\nu_0} u^\dagger$). (In Lean, with a bound G on the source norm: $f = S_{\nu_0} T_0^a g_0$ with $\|g_0\| \leq G$.)

Chapter 3

Basic lemmas of the setting

The lemmas of Section 2 of the paper: finiteness of the amplitude moments on the domain of the free energy (`lem:moment`), the Gibbs variational principle relative to a reference measure (`lem:gibbs-variational`), and the properties of the synthesis operator, its adjoint, the kernel operator, the resolvent and the regularized ridgelet transform (`lem:operators`).

3.1 Second moments on the domain

Theorem 166. *For $\nu_0 \in \mathcal{P}(Z)$ and $f\colon \mathbb{R} \to \mathbb{R}$ measurable, $\int_\Theta f(a)\,\mu_U(\mathrm{d}a\,\mathrm{d}z) = \int_{\mathbb{R}} f(a)\,\mathcal{N}(0,\beta/\lambda)(\mathrm{d}a)$.*

Proof. □

Theorem 167. *Let $\lambda, \beta > 0$, $\nu_0 \in \mathcal{P}(Z)$ and $2s\beta/\lambda < 1$. Then $a \mapsto e^{sa^2}$ is μ_U-integrable and $\int_\Theta e^{sa^2}\,\mu_U(\mathrm{d}\theta) = (1-2s\beta/\lambda)^{-1/2}$.*

Proof. □

Theorem 168. *Let $\lambda, \beta > 0$, $\nu_0 \in \mathcal{P}(Z)$, and let $\rho \in \mathcal{D}$, i.e. $\rho \in \mathcal{P}(\Theta)$ with $\mathrm{KL}(\rho\|\mu_U) < \infty$. For every $\eta \in (0,1)$,*

$$\int_\Theta a^2\,\rho(\mathrm{d}a\,\mathrm{d}z) \le \frac{2\beta}{\eta\lambda}\Big[\mathrm{KL}(\rho\|\mu_U) + \tfrac{1}{2}\log\tfrac{1}{1-\eta}\Big].$$

Only $|\varphi_z| \le 1$ of (A3) and $\nu_0 \in \mathcal{P}(Z)$ of (A4), (A5) enter.

Proof. Step 1: the test function ‘g(θ) = c a²‘ with ‘c = ηλ/(2β)‘, so that ‘2 c β/λ = η < 1‘.
Step 2: the Gaussian exponential moment ‘∫ e^g dμ_U = (1-η)^-1/2‘.
Step 3: ‘g ∈ L¹(ρ)‘ (Lebesgue-integral Donsker–Varadhan), hence ‘a² ∈ L¹(ρ)‘.
Step 4: the Donsker–Varadhan inequality ‘c ∫ a² dρ ≤ KL + log (1-η)^-1/2‘.
Step 5: ‘log (1-η)^-1/2 = ½ log (1/(1-η))‘, and divide by ‘c‘. □

Theorem 169. *Under the hypotheses of Lemma 168, taking $\eta = 1/2$, $\int_\Theta a^2\,\mathrm{d}\rho \le \frac{4\beta}{\lambda}[\mathrm{KL}(\rho\|\mu_U) + \frac{1}{2}\log 2]$.*

Proof. □

Theorem 170. *Under the hypotheses of Lemma 168, $\int_\Theta a^2\,\mathrm{d}\rho < \infty$, i.e. $a \mapsto a^2$ is ρ-integrable.*

Proof. The test function ‘g(θ) = (λ/(4β)) a²‘ of the case ‘η = 1/2‘. □

Theorem 171. *If ρ is a finite measure on Θ with $\int a^2 \, \mathrm{d}\rho < \infty$, then $\int |a| \, \mathrm{d}\rho < \infty$ (since $|a| \leq a^2 + 1$).*

Proof. ◻

Theorem 172. *Under the hypotheses of Lemma 168, $\int |a| \, \mathrm{d}\rho < \infty$ and $\int a^2 \, \mathrm{d}\rho < \infty$.*

Proof. ◻

Theorem 173. *For $\rho \in \mathcal{P}(\Theta)$ with $\int a^2 \, \mathrm{d}\rho < \infty$, $\int |a| \, \mathrm{d}\rho \leq \left(\int a^2 \, \mathrm{d}\rho\right)^{1/2}$ (Cauchy–Schwarz).*

Proof. 'Var[|a|] = ∫ a² − (∫ |a|)² ≥ 0'. ◻

Theorem 174. *Under the hypotheses of Lemma 168, $\int |a| \, \mathrm{d}\rho \leq \left(\int a^2 \, \mathrm{d}\rho\right)^{1/2} < \infty$.*

Proof. ◻

Theorem 175. *Under the hypotheses of Lemma 168, for every $x \in \mathcal{X}$, $|F_\rho(x)| \leq \int |a| \, \mathrm{d}\rho \leq \left(\int a^2 \, \mathrm{d}\rho\right)^{1/2}$.*

Proof. ◻

3.2 Gibbs variational principle

Theorem 176. *If $\int e^{-W/\beta} \, \mathrm{d}\mu < \infty$ then $\hat{\mu}_W = Z_W^{-1} e^{-W/\beta} \mu$ is the exponential tilt of μ by $-W/\beta$ (Mathlib's 'Measure.tilted'), i.e. the measure with density $e^{-W/\beta} / \int e^{-W/\beta} \, \mathrm{d}\mu$ with respect to μ.*

Proof. ◻

Theorem 177. *If $\int e^{-W/\beta} \, \mathrm{d}\mu < \infty$ then* $$\hat{\mu}_W(\mathrm{d}\theta) = \frac{e^{-W(\theta)/\beta}}{\int e^{-W/\beta} \, \mathrm{d}\mu} \mu(\mathrm{d}\theta).$$

Proof. ◻

Theorem 178. *If $\mu \neq 0$ and $Z_W = \int e^{-W/\beta} \, \mathrm{d}\mu < \infty$ then $\hat{\mu}_W$ is a probability measure.*

Proof. ◻

Theorem 179. *If $Z_W < \infty$ then $\hat{\mu}_W \ll \mu$.*

Proof. ◻

Theorem 180. *If $Z_W < \infty$ then $\mu \ll \hat{\mu}_W$, so $\hat{\mu}_W$ and μ are equivalent.*

Proof. ◻

Theorem 181. *Let μ be σ-finite, $\beta > 0$, W measurable with $Z_W < \infty$, and let $\rho \in \mathcal{P}(\Theta)$ with $\rho \ll \mu$ and $\int |W| \, \mathrm{d}\rho < \infty$. Then $\mathrm{KL}(\rho \| \hat{\mu}_W) < \infty$ if and only if $\mathrm{Ent}_\mu(\rho) = \int \log \frac{\mathrm{d}\rho}{\mathrm{d}\mu} \, \mathrm{d}\rho$ is finite, i.e. $\log \frac{\mathrm{d}\rho}{\mathrm{d}\mu} \in L^1(\rho)$.*

Proof. ◻

Theorem 182. *Under the hypotheses of Lemma 181, if $\mathrm{Ent}_\mu(\rho)$ is finite then $\mathrm{KL}(\rho \| \hat{\mu}_W) = \mathrm{Ent}_\mu(\rho) + \frac{1}{\beta} \int W \, \mathrm{d}\rho + \log Z_W$.*

Proof. 'KL(ρ‖μ̂_W) = ∫ llr ρ μ̂_W dρ' for probability measures, and Mathlib's computation of the log-likelihood ratio with respect to a tilted measure. ◻

Theorem 183. *Let μ be a σ-finite measure on Θ, $\beta > 0$, $W\colon \Theta \to \mathbb{R}$ measurable with $Z_W := \int_\Theta e^{-W/\beta}\,\mathrm{d}\mu < \infty$, and $\hat\mu_W(\mathrm{d}\theta) := Z_W^{-1} e^{-W(\theta)/\beta}\mu(\mathrm{d}\theta)$. If $\rho \in \mathcal{P}(\Theta)$, $\rho \ll \mu$, $\int |W|\,\mathrm{d}\rho < \infty$ and $\operatorname{Ent}_\mu(\rho) = \int \log\frac{\mathrm{d}\rho}{\mathrm{d}\mu}\,\mathrm{d}\rho$ is finite, then*

$$\int W\,\mathrm{d}\rho + \beta\,\operatorname{Ent}_\mu(\rho) = \beta\,\mathrm{KL}(\rho\|\hat\mu_W) - \beta\log Z_W.$$

(The paper takes μ to be Lebesgue measure; the case $\operatorname{Ent}_\mu(\rho) = +\infty$ is Lemma 181.)

Proof. □

Theorem 184. *Under the hypotheses of Lemma 183, $\int W\,\mathrm{d}\rho + \beta\,\operatorname{Ent}_\mu(\rho) \geq -\beta\log Z_W$, with equality if and only if $\rho = \hat\mu_W$ (Lemma 185).*

Proof. □

Theorem 185. *Under the hypotheses of Lemma 183, $\int W\,\mathrm{d}\rho + \beta\,\operatorname{Ent}_\mu(\rho) = -\beta\log Z_W$ if and only if $\rho = \hat\mu_W$.*

Proof. □

Theorem 186. *Let $\mu \in \mathcal{P}(\Theta)$, $\beta > 0$, W measurable with $Z_W < \infty$, and $\rho \in \mathcal{P}(\Theta)$ with $\rho \ll \mu$ and $\int |W|\,\mathrm{d}\rho < \infty$. Then $\mathrm{KL}(\rho\|\hat\mu_W) < \infty$ if and only if $\mathrm{KL}(\rho\|\mu) < \infty$.*

Proof. □

Theorem 187. *Let $\mu \in \mathcal{P}(\Theta)$, $\beta > 0$, W measurable with $Z_W < \infty$, and $\rho \in \mathcal{P}(\Theta)$ with $\rho \ll \mu$, $\int |W|\,\mathrm{d}\rho < \infty$ and $\mathrm{KL}(\rho\|\mu) < \infty$. Then $\mathrm{KL}(\rho\|\hat\mu_W) = \mathrm{KL}(\rho\|\mu) + \frac{1}{\beta}\int W\,\mathrm{d}\rho + \log Z_W$, i.e. $\int W\,\mathrm{d}\rho + \beta\,\mathrm{KL}(\rho\|\mu) = \beta\,\mathrm{KL}(\rho\|\hat\mu_W) - \beta\log Z_W$.*

Proof. □

Theorem 188. *Under the hypotheses of Lemma 187, $\int W\,\mathrm{d}\rho + \beta\,\mathrm{KL}(\rho\|\mu) \geq -\beta\log Z_W$, with equality if and only if $\rho = \hat\mu_W$.*

Proof. □

3.3 Synthesis operator, adjoint, kernel and regularized ridgelet transform

Theorem 189. *For $r \in L^2(P_X)$ and every $z \in Z$, $|S^* r(z)| \leq \|\varphi_z\|_{L^2(P_X)}\|r\|_{L^2(P_X)} \leq \|r\|_{L^2(P_X)}$.*

Proof. □

Theorem 190. *For $r \in L^2(P_X)$ and every $\nu \in \mathcal{P}(Z)$, $S^* r \in L^2(\nu)$: it is measurable (Fubini) and bounded by $\|r\|_{L^2(P_X)}$.*

Proof. □

Theorem 191. *Let Z be a metric space and assume $|\varphi_z(x) - \varphi_{z'}(x)| \leq L\,d(z,z')$ for all x, z, z'. Then for $r \in L^2(P_X)$, $|S^* r(z) - S^* r(z')| \leq L\,\|r\|_{L^2(P_X)}\,d(z,z')$; under (A1) and (A3) this is* (**??**) *with $L = \sqrt{R_X^2 + 1}$.*

Proof. Step 1: the difference is the integral of ‘(φ_z − φ_z’) r‘.
Step 2: ‘L · dist z z’ $\geq$ 0‘ (the input space is nonempty since ‘P‘ is a probability measure).
Step 3: bound the integrand by ‘L · dist z z’ · |r|‘ and use ‘‖r‖_L¹ $\leq$ ‖r‖_L²‘. □

Theorem 192. *For every $\nu \in \mathcal{P}(Z)$ and $r \in L^2(P_X)$ the adjoint of the synthesis operator is the analysis map: $(S_\nu^* r)(z) = \langle \varphi_z, r\rangle_{L^2(P_X)} = (S^* r)(z)$ for ν-a.e. z. In particular $S_\nu^* r$ does not depend on ν (as a function).*

Proof. □

Theorem 193. *$K_\nu = S_\nu S_\nu^*$ is a self-adjoint positive semidefinite operator on $L^2(P_X)$: $\langle K_\nu r, r\rangle_{L^2(P_X)} = \|S_\nu^* r\|^2_{L^2(\nu)} \geq 0$ for every r, and $\|K_\nu\| \leq \|S_\nu\|\|S_\nu^*\| \leq 1$.*

Proof. □

Theorem 194. *For $r \in L^2(P_X)$, P_X-a.e. in x, $(K_\nu r)(x) = \int_Z \varphi_z(x)\,(S^* r)(z)\,\nu(\,\mathrm{d}z) = \int_Z \varphi_z(x)\langle \varphi_z, r\rangle_{L^2(P_X)}\,\nu(\,\mathrm{d}z)$, i.e. K_ν is the integral operator with kernel $k_\nu(x, x') = \int \varphi_z(x)\varphi_z(x')\,\nu(\,\mathrm{d}z)$.*

Proof. □

Theorem 195. *For $\lambda > 0$ the operator $K_\nu + \lambda$ is a bijection of $L^2(P_X)$ with bounded inverse: $(K_\nu + \lambda)(K_\nu+\lambda)^{-1} = \mathrm{id}$ and $(K_\nu+\lambda)^{-1}(K_\nu+\lambda) = \mathrm{id}$, where $(K_\nu+\lambda)^{-1}$ is ‘kernelResolvent‘. (Proof: $K_\nu \geq 0$ gives $\langle (K_\nu + \lambda)r, r\rangle \geq \lambda\|r\|^2$, so $K_\nu + \lambda$ is injective with closed range whose orthogonal complement is trivial.)*

Proof. □

Theorem 196. *For $\lambda > 0$, $\|(K_\nu + \lambda)^{-1}\| \leq 1/\lambda$: for $g = (K_\nu + \lambda)^{-1} f$, $\lambda\|g\|^2 \leq \langle (K_\nu + \lambda)g, g\rangle = \langle f, g\rangle \leq \|f\|\|g\|$.*

Proof. □

Theorem 197. *For $\lambda > 0$, $\|\lambda(K_\nu + \lambda)^{-1}\| \leq 1$ and $\|K_\nu(K_\nu + \lambda)^{-1}\| \leq 1$. (Proof without spectral calculus: for $g = (K_\nu + \lambda)^{-1} f$, $\|f\|^2 = \|(K_\nu + \lambda)g\|^2 = \|K_\nu g\|^2 + 2\lambda\langle K_\nu g, g\rangle + \lambda^2\|g\|^2$ dominates both $\lambda^2\|g\|^2$ and $\|K_\nu g\|^2$ since $\langle K_\nu g, g\rangle \geq 0$.)*

Proof. □

Theorem 198. *For $\lambda > 0$ and $f \in L^2(P_X)$, $\|R_{\lambda,\nu} f\|_{L^2(\nu)} \leq \|f\|_{L^2(P_X)}/(2\sqrt{\lambda})$. (Proof without the spectral measure: with $g = (K_\nu+\lambda)^{-1} f$, $\|S_\nu^* g\|^2 = \langle K_\nu g, g\rangle$ and $\|f\|^2 - 4\lambda\langle K_\nu g, g\rangle = \|(K_\nu - \lambda)g\|^2 \geq 0$.)*

Proof. □

Theorem 199. *For $\lambda > 0$, $f \in L^2(P_X)$ and every $z \in Z$, $|R_{\lambda,\nu} f(z)| \leq \|(K_\nu + \lambda)^{-1} f\|_{L^2(P_X)} \leq \|f\|_{L^2(P_X)}/\lambda$.*

Proof. □

Theorem 200. *For $\lambda > 0$ and $f \in L^2(P_X)$, $u = S_\nu^*(K_\nu + \lambda)^{-1} f$ solves $(S_\nu^* S_\nu + \lambda)u = S_\nu^* f$; i.e. the push-through identity $S_\nu^*(K_\nu + \lambda)^{-1} f = (S_\nu^* S_\nu + \lambda)^{-1} S_\nu^* f$ holds.*

Proof. □

Theorem 201. *For $\lambda > 0$ and $f \in L^2(P_X)$, $u_0 = S_\nu^*(K_\nu + \lambda)^{-1} f$ is the unique minimizer over $L^2(\nu)$ of the Tikhonov functional $J(u) = \frac{1}{2}\|S_\nu u - f\|^2_{L^2(P_X)} + \frac{\lambda}{2}\|u\|^2_{L^2(\nu)}$: $J(u_0) \leq J(u)$ for all u, with equality only for $u = u_0$. (Proof: by the normal equation, $J(u) - J(u_0) = \frac{1}{2}\|S_\nu(u - u_0)\|^2 + \frac{\lambda}{2}\|u - u_0\|^2$.)*

Proof. □

3.4 Characterization of the minimizer of the free energy

Theorem 202. *For ρ with $\int |a|\,\mathrm{d}\rho < \infty$, $L(\rho) = \frac{1}{2}\|r_\rho\|^2_{L^2(P_X)}$ with $r_\rho = F_\rho - f \in L^2(P_X)$.*

Proof. □

Theorem 203. *For ρ with $\int |a|\,\mathrm{d}\rho < \infty$ and $g \in L^2(P_X)$, $\int_\Theta a\,(S^*g)(z)\,\rho(\,\mathrm{d}\theta) = \langle F_\rho, g\rangle_{L^2(P_X)}$ (Fubini, since $\iint |a||\varphi_z(x)||g(x)|\,\rho(\,\mathrm{d}\theta)P_X(\,\mathrm{d}x) \le \int |a|\,\mathrm{d}\rho\,\|g\|_{L^1(P_X)} < \infty$).*

Proof. Step 1: the integrand ‘a φ_z(x) g(x)‘ is integrable on ‘P ⊗ ρ‘, being bounded by ‘|g(x)| |a|‘.
Step 2: write the inner product as an integral and swap the order of integration. □

Theorem 204. *For $\rho, \rho' \in \mathcal{D}$ (indeed for any ρ, ρ' with $\int |a|\,\mathrm{d}\rho, \int |a|\,\mathrm{d}\rho' < \infty$), with $s_\rho := S^*r_\rho$,*

$$L(\rho') - L(\rho) = \int_\Theta a\,s_\rho(z)\ \mathrm{d}(\rho' - \rho)(\theta) + \tfrac{1}{2}\|F_{\rho'} - F_\rho\|^2_{L^2(P_X)}.$$

(Proof: $r_{\rho'} = r_\rho + (F_{\rho'} - F_\rho)$, expand the square, and use Lemma 203.)

Proof. ‘r_ρ’ = r_ρ + (F_ρ’ − F_ρ)‘ and ‘‖r + d‖² = ‖r‖² + 2⟪r, d⟫ + ‖d‖²‘. □

Theorem 205. *Under the hypotheses of Lemma 204, $L(\rho') \ge L(\rho) + \int a\,s_\rho(z)\ \mathrm{d}(\rho' - \rho)$, with equality if and only if $F_{\rho'} = F_\rho$ in $L^2(P_X)$. In particular L is convex on $\mathcal{D}$.*

Proof. □

Theorem 206. *Let $s\colon Z \to \mathbb{R}$ be measurable with $|s| \le C$ and $W(\theta) := a\,s(z)$. Then $\int_\Theta e^{-W/\beta}\,\mathrm{d}\mu_U < \infty$, since $e^{-as(z)/\beta} \le e^{C|a|/\beta}$ and the a-marginal of μ_U is Gaussian.*

Proof. □

Theorem 207. *Let s be measurable with $|s| \le C$ and $W(\theta) = a\,s(z)$. Then $\hat{\mu}_W = Z_W^{-1}e^{-W/\beta}\mu_U$ is a probability measure with $\mathrm{KL}(\hat{\mu}_W\|\mu_U) < \infty$, since $\log\frac{\mathrm{d}\hat{\mu}_W}{\mathrm{d}\mu_U} = -\frac{as(z)}{\beta} - \log Z_W$ is $\hat{\mu}_W$-integrable.*

Proof. □

Theorem 208. *Let s be measurable with $|s| \le C$, $W(\theta) = a\,s(z)$ and $\hat{\mu} := \hat{\mu}_W$ (relative to μ_U). For every $\rho \in \mathcal{D}$, $\mathrm{KL}(\rho\|\hat{\mu}) < \infty$ and*

$$\mathcal{F}(\rho) = L(\rho) - \int_\Theta a\,s(z)\,\mathrm{d}\rho + \beta\ \mathrm{KL}(\rho\|\hat{\mu}) - \beta\log Z_W.$$

(Chain rule $\mathrm{KL}(\rho\|\mu_U) = \mathrm{KL}(\rho\|\hat{\mu}) + \int \log\frac{\mathrm{d}\hat{\mu}}{\mathrm{d}\mu_U}\,\mathrm{d}\rho$, legitimate since $\int |a||s|\,\mathrm{d}\rho \le C\int |a|\,\mathrm{d}\rho < \infty$ by Lemma 168.)

Proof. □

Theorem 209. *Let $\rho_0 \in \mathcal{D}$ satisfy the self-consistency equation $\rho_0 = \hat{\mu}_{W_{\rho_0}}$, $W_{\rho_0}(\theta) = a\,(S^*r_{\rho_0})(z)$. Then for every $\rho \in \mathcal{D}$, $\mathrm{KL}(\rho\|\rho_0) < \infty$ and $\mathcal{F}(\rho) - \mathcal{F}(\rho_0) = \frac{1}{2}\|F_\rho - F_{\rho_0}\|^2_{L^2(P_X)} + \beta\,\mathrm{KL}(\rho\|\rho_0)$. (Lemmas 208 and 204 with $\hat{\mu} = \rho_0$.)*

Proof. Step 1: the decomposition of ‘$\mathcal{F}$‘ relative to ‘μhat = ρ_0‘ at ‘ρ‘ and at ‘ρ_0‘.
Step 2: the expansion of ‘L‘ around ‘ρ_0‘. □

Theorem 210. *If $\rho_0 \in \mathcal{D}$ solves the self-consistency equation $\rho_0 = \hat{\mu}_{W_{\rho_0}}$ with $W_{\rho_0}(\theta) = a\,(S^* r_{\rho_0})(z)$, then $\mathcal{F}(\rho) \geq \mathcal{F}(\rho_0) + \beta\,\mathrm{KL}(\rho\|\rho_0) \geq \mathcal{F}(\rho_0)$ for every $\rho \in \mathcal{D}$, so ρ_0 is a minimizer of $\mathcal{F}$ on $\mathcal{D}$.*

Proof. □

Theorem 211. *Let $\rho^* \in \mathcal{D}$ be a minimizer of $\mathcal{F}$ on $\mathcal{D}$, $s^* := S^* r_{\rho^*}$ and $W^*(\theta) := a\,s^*(z)$. Then $Z_{W^*} < \infty$ and $\rho^* = \hat{\mu}_{W^*}$, i.e. $\rho^*(\mathrm{d}\theta) = Z_{W^*}^{-1} e^{-a s^*(z)/\beta}\,\mu_U(\mathrm{d}\theta)$; with $\mu_U \propto e^{-U/\beta}\,\mathrm{d}\theta$ this is $\rho^*(\mathrm{d}a\,\mathrm{d}z) \propto \exp(-\frac{a s^*(z) + \frac{\lambda}{2}a^2 + V(z)}{\beta})\,\mathrm{d}a\,\mathrm{d}z$. (Proof: with $\hat{\mu} := \hat{\mu}_{W^*} \in \mathcal{D}$ and $\rho_\varepsilon := (1-\varepsilon)\rho^* + \varepsilon\hat{\mu}$, Lemmas 208, 204 and 904 give $0 \leq \mathcal{F}(\rho_\varepsilon) - \mathcal{F}(\rho^*) \leq \frac{\varepsilon^2}{2}\|F_{\hat{\mu}} - F_{\rho^*}\|^2 - \varepsilon\beta\,\mathrm{KL}(\rho^*\|\hat{\mu})$; let $\varepsilon \downarrow 0$.)*

Proof. Step 1: ‘μhat = gibbs μ_U β (a s(z))‘ is a probability measure in the domain.
Step 2: the decomposition at ‘ρ_0‘, with ‘K = KL(ρ_0‖μhat)‘ and ‘D = ‖F_μhat − F_ρ_0‖²‘.
Step 3: for ‘ε ∈ (0,1)‘ and the mixture ‘ρ_ε = (1−ε) ρ_0 + ε μhat ∈ $\mathcal{D}$‘, minimality, the decomposition, the expansion and the convexity of ‘KL‘ give ‘β K ≤ ε D / 2‘.
Step 4: letting ‘ε ↓ 0‘ gives ‘K = 0‘, hence ‘ρ_0 = μhat‘. □

Theorem 212. *Let $\rho^* \in \mathcal{D}$ be a minimizer of $\mathcal{F}$ on $\mathcal{D}$. For every $\rho \in \mathcal{D}$, $\mathrm{KL}(\rho\|\rho^*) < \infty$ and*

$$\mathcal{F}(\rho) - \mathcal{F}(\rho^*) = \tfrac{1}{2}\|F_\rho - F_{\rho^*}\|^2_{L^2(P_X)} + \beta\ \mathrm{KL}(\rho\|\rho^*)\ \geq\ \beta\ \mathrm{KL}(\rho\|\rho^*).$$

(Lemma 209 with $\rho_0 = \rho^*$, which is self-consistent by Lemma 211.)

Proof. □

Theorem 213. *The minimizer of $\mathcal{F}$ on $\mathcal{D}$ is unique: if $\rho_1, \rho_2 \in \mathcal{D}$ both minimize $\mathcal{F}$ then $\rho_1 = \rho_2$. (By Lemma 212, $0 \geq \mathcal{F}(\rho_2) - \mathcal{F}(\rho_1) \geq \beta\,\mathrm{KL}(\rho_2\|\rho_1)$, so $\mathrm{KL}(\rho_2\|\rho_1) = 0$.)*

Proof. □

3.5 Existence of the minimizer

Theorem 214. *For $T > 0$ and $a \in \mathbb{R}$, $|a - \max(-T, \min(T, a))| \leq a^2/T$.*

Proof. □

Theorem 215. *Let Z be a topological space with its Borel σ-algebra, let $z \mapsto \varphi_z(x)$ be continuous for every x, and let $\rho_n \to \rho$ weakly in $\mathcal{P}(\Theta)$ with $\sup_n \int a^2\,\mathrm{d}\rho_n \leq C$ and $\int a^2\,\mathrm{d}\rho \leq C$. Then $F_{\rho_n}(x) \to F_\rho(x)$ for every $x \in \mathcal{X}$. Proof: for $T > 0$ the function $\theta \mapsto \max(-T, \min(T, a))\,\varphi_z(x)$ is bounded and continuous, and replacing a by its clamped value changes $F_\rho(x)$ by at most $\int a^2\,\mathrm{d}\rho/T \leq C/T$ (Lemma 214).*

Proof. Step 1: the truncation level ‘T = 4 max(C,1)/ε‘ and the bounded continuous test function ‘h_T(θ) = max (−T) (min T a) φ_z(x)‘.
Step 2: for every probability measure ‘σ‘ with ‘∫ a² dσ ≤ C‘, ‘|F_σ(x) − ∫ h_T dσ| ≤ ∫ a²/T dσ ≤ C_1/T = ε/4‘.
Step 3: weak convergence gives ‘|∫ h_T dρ_n − ∫ h_T dρ| < ε/2‘ for ‘n‘ large, and the triangle inequality concludes. □

Theorem 216. *Under the hypotheses of Lemma 215, for $P_X \in \mathcal{P}(\mathcal{X})$ and $f \in L^2(P_X)$, $L(\rho_n) \to L(\rho)$. Proof: $F_{\rho_n}(x) \to F_\rho(x)$ pointwise and $|F_{\rho_n}| \leq \sqrt{C}$, so $(F_{\rho_n} - f)^2 \leq 2C + 2f^2 \in L^1(P_X)$ and dominated convergence applies.*

Proof. Step 1: the uniform bound ‘$F\_\sigma(x)^2 \leq \int a^2\, d\sigma \leq C$‘.

Step 2: dominated convergence with the bound ‘$2C + 2f^2$‘. □

Theorem 217. *Let Z be a Polish space with its Borel σ-algebra, let $z \mapsto \varphi_z(x)$ be continuous for every x (both hold under (A3)–(A5)), and let $f \in L^2(P_X)$. Then the free energy $\mathcal{F}$ has a minimizer on $\mathcal{D}$: there is $\rho^* \in \mathcal{D}$ with $\mathcal{F}(\rho^*) \leq \mathcal{F}(\rho)$ for all $\rho \in \mathcal{D}$. (Direct method: tightness of the sublevel sets of* $\mathrm{KL}(\cdot\|\mu_U)$*, lower semicontinuity of* KL*, Prokhorov, and continuity of L along weakly convergent sequences with uniformly bounded second amplitude moments.)*

Proof. Step 1: the free energy is nonnegative, the domain contains ‘μ_U‘; let ‘m‘ be the infimum of ‘$\mathcal{F}$‘ on ‘$\mathcal{D}$‘.

Step 2: a minimizing sequence ‘$\rho_n \in \mathcal{D}$‘ with ‘$\mathcal{F}(\rho_n) < m + 1/(n+1)$‘, so ‘$\mathcal{F}(\rho_n) \to m$‘.

Step 3: the uniform bounds ‘$\mathrm{KL}(\rho_n\|\mu\_U) \leq C := (m + 1)/\beta$‘ (since ‘$\mathcal{F} \geq \beta\, \mathrm{KL}$‘) and ‘$\int a^2\, d\rho_n \leq C' := (4\beta/\lambda)(C + \tfrac{1}{2} \log 2)$‘ (‘lem:moment‘).

Step 4: tightness of ‘ρ_n‘ and Prokhorov’s theorem: a subsequence ‘$\rho\_\varphi\ n$‘ converges weakly to some ‘$\rho\mathrm{lim}$‘.

Step 5: lower semicontinuity, ‘$\mathrm{KL}(\rho\mathrm{lim}\|\mu\_U) \leq \liminf \mathrm{KL}(\rho\_\varphi\ n\|\mu\_U) \leq C$‘: ‘$\rho\mathrm{lim} \in \mathcal{D}$‘.

Step 6: continuity of the risk along the subsequence, ‘$L(\rho\_\varphi\ n) \to L(\rho\mathrm{lim})$‘.

Step 7: ‘$\mathrm{KL}(\rho\_\varphi\ n\|\mu\_U) = (\mathcal{F}(\rho\_\varphi\ n) - L(\rho\_\varphi\ n))/\beta \to k := (m - L(\rho\mathrm{lim}))/\beta$‘, and by lower semicontinuity ‘$\mathrm{KL}(\rho\mathrm{lim}\|\mu\_U) \leq k$‘.

Step 8: ‘$\mathcal{F}(\rho\mathrm{lim}) = L(\rho\mathrm{lim}) + \beta\, \mathrm{KL}(\rho\mathrm{lim}\|\mu\_U) \leq L(\rho\mathrm{lim}) + \beta\, k = m \leq \mathcal{F}(\rho')$‘ for all ‘$\rho' \in \mathcal{D}$‘. □

Chapter 4

Structure of the minimizers: the null-space effect

Section 3 of the paper: the structure of the Gibbs minimizer of the free energy (`thm:m1`: Gaussian conditional amplitudes whose mean is the regularized ridgelet transform with respect to the learned hidden marginal, kernel ridge regression of the learned kernel), the structure of the stationary points of the noiseless finite-width gradient flow with weight decay (`thm:m1prime`) and the fixed-feature flow without amplitude regularization, whose endpoint keeps the null-space component of the initialization (`prop:no-regularization`, `cor:necessity-amplitude`).

4.1 Gibbs minimizer of the free energy

Theorem 218. *Let s be measurable with $|s| \le C$, $W_s(\theta) = a\,s(z)$ and $\rho_s := \hat{\mu}_{W_s}$ relative to μ_U. Then $\rho_s(\,\mathrm{d}a\,\mathrm{d}z) = \nu[s](\,\mathrm{d}z)\,\mathcal{N}(-s(z)/\lambda, \beta/\lambda)(\,\mathrm{d}a)$ with $\nu[s](\,\mathrm{d}z) = Z_*^{-1} e^{s(z)^2/(2\lambda\beta)}\nu_0(\,\mathrm{d}z)$, $Z_* = \int e^{s^2/(2\lambda\beta)}\,\mathrm{d}\nu_0$. (Completing the square, $as + \frac{\lambda}{2}a^2 = \frac{\lambda}{2}(a + s/\lambda)^2 - \frac{s^2}{2\lambda}$.)*

Proof. □

Theorem 219. *Under the hypotheses of Lemma 218, $\nu_{\rho_s} = \nu[s]$, i.e. $\nu_{\rho_s}(\,\mathrm{d}z) = Z_*^{-1} e^{s(z)^2/(2\lambda\beta)}\nu_0(\,\mathrm{d}z)$.*

Proof. □

Theorem 220. *Under the hypotheses of Lemma 218, $m_{\rho_s}(z) = -s(z)/\lambda$ for ν_{ρ_s}-a.e. z.*

Proof. □

Theorem 221. *Under the hypotheses of Lemma 218, $F_{\rho_s}(x) = \int_Z (-s(z)/\lambda)\varphi_z(x)\,\nu_{\rho_s}(\,\mathrm{d}z) = (S_{\nu_{\rho_s}}(-s/\lambda))(x)$ for every x.*

Proof. □

Theorem 222. *Let s be measurable with $|s| \le C$ and $c := C^2/(2\lambda\beta)$. Then $\nu[s] = w\,\nu_0$ with $w = Z_*^{-1}e^{s^2/(2\lambda\beta)}$, $1 \le Z_* \le e^c$, hence $e^{-c} \le w \le e^c$ everywhere; in particular $\nu[s]$ and ν_0 are mutually absolutely continuous.*

Proof. □

Theorem 223. *Under the hypotheses of Lemma 218, $\int |a|\,\mathrm{d}\rho_s \le \int |s/\lambda|\,\mathrm{d}\nu_{\rho_s} + \sqrt{2\beta/(\pi\lambda)} \le C/\lambda + \sqrt{2\beta/(\pi\lambda)}$ (the Gaussian bound $\mathbb{E}|X| \le |\mu| + \sigma\sqrt{2/\pi}$ for $X \sim \mathcal{N}(\mu, \sigma^2)$).*

Proof. Step 1: the inner Gaussian moment bound.
Step 2: integrate against the hidden marginal. □

Theorem 224. *Under the hypotheses of Lemma 218, $\int a^2\,\mathrm{d}\rho_s = \int (s/\lambda)^2\,\mathrm{d}\nu_{\rho_s} + \beta/\lambda$ (the Gaussian identity $\mathbb{E}X^2 = \mu^2 + \sigma^2$).*

Proof. □

Theorem 225. *Let ρ^* be the minimizer of $\mathcal{F}$ on $\mathcal{D}$, $s^* := S^* r_{\rho^*}$ and $\nu^* := \nu_{\rho^*}$. Then $\rho^*(\mathrm{d}a\,\mathrm{d}z) = \nu^*(\mathrm{d}z)\,\mathcal{N}(-s^*(z)/\lambda, \beta/\lambda)(\mathrm{d}a)$.*

Proof. □

Theorem 226. *With the notation of Theorem 225, $\nu^*(\mathrm{d}z) = Z_*^{-1}\exp(\frac{s^*(z)^2}{2\lambda\beta})\nu_0(\mathrm{d}z)$ with $Z_* = \int_Z \exp(\frac{s^*(z)^2}{2\lambda\beta})\nu_0(\mathrm{d}z) < \infty$; with $\nu_0 = Z_0^{-1}e^{-V/\beta}\,\mathrm{d}z$ this is $\nu^*(\mathrm{d}z) \propto \exp(\frac{s^*(z)^2}{2\lambda\beta} - \frac{V(z)}{\beta})\,\mathrm{d}z$.*

Proof. □

Theorem 227. *The measures ν^* and ν_0 are mutually absolutely continuous.*

Proof. □

Theorem 228. *$m_{\rho^*}(z) = -s^*(z)/\lambda$ for ν^*-a.e. z (the conditional law of a given z is $\mathcal{N}(-s^*(z)/\lambda, \beta/\lambda)$).*

Proof. □

Theorem 229. *$\Pi\rho^* = m_{\rho^*}\,\nu^*$ with $m_{\rho^*} = -s^*/\lambda$, i.e. $\Pi\rho^*(A) = \int_A (-s^*(z)/\lambda)\nu^*(\mathrm{d}z)$ for measurable $A \subseteq Z$.*

Proof. Step 1: transport the set integral through the disintegration.
Step 2: the inner integral is the Gaussian mean '−s(z)/λ'. □

Theorem 230. *In $L^2(\nu^*)$, $m_{\rho^*} = -\lambda^{-1}S_{\nu^*}^* r^* \in \operatorname{ran} S_{\nu^*}^* \subseteq (\ker S_{\nu^*})^\perp$.*

Proof. □

Theorem 231. *With $K^* := K_{\nu^*}$,*

$$r^* = -\lambda(K^*{+}\lambda)^{-1}f, \qquad F_{\rho^*} = K^*(K^*{+}\lambda)^{-1}f, \qquad m_{\rho^*} = S^*(K^*{+}\lambda)^{-1}f = R_{\lambda,\nu^*}f\ (\nu^*\text{-a.e.}), \qquad s^* = -\lambda R_{\lambda,\nu^*}f.$$

(Proof: $F_{\rho^*} = S_{\nu^*}m_{\rho^*} = -\lambda^{-1}K^*r^*$ by Lemma 221 and Theorem 230, so $(K^*+\lambda)r^* = -\lambda f$, and $K^*+\lambda$ is invertible.)

Proof. Step 1: 'F_ρ* = S_ν* (−s*/λ) = −λ⁻¹ K* r*'.
Step 2: '(K* + λ) r* = −λ f', hence 'r* = −λ (K* + λ)⁻¹ f'.
Step 3: 'F_ρ* = f + r* = K* (K* + λ)⁻¹ f'.
Step 4: 's* = −λ R f' and 'm* = R f'. □

Theorem 232. *With $\|f\| := \|f\|_{L^2(P_X)}$, $\|r^*\| \le \|f\|$, $\sup_z |s^*(z)| \le \|f\|$, $|m_{\rho^*}| \le \|f\|/\lambda$ and $\|m_{\rho^*}\|_{L^2(\nu^*)} \le \|f\|/(2\sqrt{\lambda})$.*

Proof. '‖r*‖ = ‖λ (K* + λ)⁻¹ f‖ ≤ ‖f‖' by 'lem:operators' (5).
'‖m*‖_L²(ν*) = ‖S_ν** (K* + λ)⁻¹ f‖ ≤ ‖f‖/(2 √λ)' by 'lem:operators' (4). □

Theorem 233. *$\int |a|\,\mathrm{d}\rho^* \le \frac{\|f\|}{\lambda} + \sqrt{\frac{2\beta}{\pi\lambda}}$ and $\int a^2\,\mathrm{d}\rho^* = \|m_{\rho^*}\|^2_{L^2(\nu^*)} + \frac{\beta}{\lambda} \le \min\{\frac{\|f\|^2}{\lambda^2}, \frac{\|f\|^2}{4\lambda}\} + \frac{\beta}{\lambda}$.*

Proof. '‖m*‖² = ∫ (s*/λ)² dν*'.
The two bounds on '‖m*‖': '‖f‖/λ' (sup bound) and '‖f‖/(2 √λ)'. □

Theorem 234. *With $c := \|f\|^2/(2\lambda\beta)$, $\frac{\mathrm{d}\nu^*}{\mathrm{d}\nu_0}(z) = \frac{Z_0}{Z_*}\exp(\frac{s^*(z)^2}{2\lambda\beta})$ and $e^{-c} \le \frac{\mathrm{d}\nu^*}{\mathrm{d}\nu_0}(z) \le e^c$ for all z.*

Proof. □

Theorem 235. *$L(\rho^*) = \frac{\lambda^2}{2}\|(K^* + \lambda)^{-1} f\|^2_{L^2(P_X)} \le \frac{1}{2}\|f\|^2$.*

Proof. □

Theorem 236. *The self-consistency equation $s = -\lambda\, S^*(K_{\nu[s]}+\lambda)^{-1} f$, $\nu[s](\,\mathrm{d}z) \propto \exp(\frac{s(z)^2}{2\lambda\beta})\nu_0(\,\mathrm{d}z)$, has exactly one bounded measurable solution, namely s^*: s^* solves it by Theorems 226 and 231, and any bounded measurable solution s satisfies $s = s^*$. (Proof: $\rho_s := \hat{\mu}_{W_s}$ with $W_s = a\, s(z)$ has $\nu_{\rho_s} = \nu[s]$, $F_{\rho_s} = K_{\nu[s]}(K_{\nu[s]} + \lambda)^{-1} f$ and $S^* r_{\rho_s} = s$, so ρ_s solves the self-consistency equation of Lemma 210 and is the minimizer.)*

Proof. Step 1: 'F_ρ_s = S_ν[s] R_λ,ν[s] f = K_ν[s] (K_ν[s] + λ)⁻¹ f'.
Step 2: 'r_ρ_s = −λ (K_ν[s] + λ)⁻¹ f' and 'S* r_ρ_s = s'.
Step 3: 'ρ_s' solves the self-consistency equation, hence is the minimizer. □

4.2 Stationary points of the gradient flow with weight decay

Theorem 237. *Under (A3), for every $x \in \mathbb{R}^m$ the map $z = (w, b) \mapsto \varphi_z(x) = \tanh(w^\top x - b)$ is differentiable on $Z = \mathbb{R}^m \times \mathbb{R}$.*

Proof. □

Theorem 238. *Under (A3) and (A6), $J_N \in C^1(\Theta^M)$; in Lean: J_N is (Fréchet) differentiable on $\Theta^M = (\mathbb{R} \times Z)^M$ whenever $z \mapsto \varphi_z(x)$ and V are differentiable.*

Proof. □

Theorem 239. *(Gradient of J_N.) Under (A3) and (A6), $J_N \in C^1(\Theta^M)$ and, with $r_i = F_\theta(x_i) - y_i$, $\partial_{a_\ell} J_N(\theta) = \frac{1}{M}[(S_N^* r)(z_\ell) + \lambda a_\ell]$, $\nabla_{z_\ell} J_N(\theta) = \frac{1}{M}[a_\ell \nabla_z (S_N^* r)(z_\ell) + \nabla V(z_\ell)]$. In Lean the two partial derivatives are packaged as the Fréchet derivative applied to a direction $v = (v_\ell^a, v_\ell^z)_\ell$: $\partial J_N(\theta)v = \frac{1}{M}\sum_\ell([(S_N^* r)(z_\ell) + \lambda a_\ell]v_\ell^a + [a_\ell\, \partial(S_N^* r)(z_\ell) + \partial V(z_\ell)]v_\ell^z)$.*

Proof. □

Theorem 240. *Let θ° be a stationary point of J_N ($\nabla J_N(\theta^\circ) = 0$), $r_i^\circ := F_{\theta^\circ}(x_i) - y_i$ and $s^\circ := S_N^* r^\circ$. Then $a_\ell^\circ = -s^\circ(z_\ell^\circ)/\lambda$ for $\ell = 1, \dots, M$.*

Proof. □

Theorem 241. *In the setting of 240, with $\rho_M^\circ = \frac{1}{M}\sum_\ell \delta_{\theta_\ell^\circ}$ and $\nu_M^\circ = \frac{1}{M}\sum_\ell \delta_{z_\ell^\circ}$, $\Pi\rho_M^\circ = \frac{1}{M}\sum_\ell a_\ell^\circ \delta_{z_\ell^\circ} = -\lambda^{-1} s^\circ\, \nu_M^\circ$; in Lean: for every measurable $A \subset Z$, $\Pi\rho_M^\circ(A) = -\lambda^{-1}\frac{1}{M}\sum_{\ell: z_\ell^\circ \in A} s^\circ(z_\ell^\circ)$.*

Proof. ☐

Theorem 242. *In the setting of 240, each z°_ℓ is a critical point of $\Psi^\circ(z) := \frac{s^\circ(z)^2}{2\lambda} - V(z)$: $\nabla\Psi^\circ(z^\circ_\ell) = \frac{s^\circ(z^\circ_\ell)\nabla s^\circ(z^\circ_\ell)}{\lambda} - \nabla V(z^\circ_\ell) = 0$.*

Proof. ☐

Theorem 243. *In the setting of 240, with $K^\circ := K_{N,\nu^\circ_M}$, the residual satisfies $(K^\circ + \lambda)r^\circ = -\lambda y$.*

Proof. ☐

Theorem 244. *In the setting of 240, with $K^\circ := K_{N,\nu^\circ_M}$, $r^\circ = -\lambda(K^\circ + \lambda)^{-1}y$.*

Proof. ☐

Theorem 245. *In the setting of 244, $(F_{\theta^\circ}(x_i))_i = y + r^\circ = K^\circ(K^\circ + \lambda)^{-1}y$.*

Proof. ☐

Theorem 246. *In the setting of 244, $a^\circ_\ell = [S^*_N(K^\circ+\lambda)^{-1}y](z^\circ_\ell)$, the value at z°_ℓ of $R^{(N)}_{\lambda,\nu^\circ_M}y := S^*_N(K^\circ+\lambda)^{-1}y$.*

Proof. ☐

Theorem 247. *In the setting of 244, for every $x \in \mathcal{X}$, $F_{\theta^\circ}(x) = \frac{1}{N}\sum_{i=1}^N k_{\nu^\circ_M}(x, x_i)\,\alpha^\circ_i$ with $\alpha^\circ := (K^\circ + \lambda)^{-1}y$ and $k_{\nu^\circ_M}(x, x') := \frac{1}{M}\sum_\ell \varphi_{z^\circ_\ell}(x)\varphi_{z^\circ_\ell}(x')$.*

Proof. ☐

Theorem 248. *In the setting of 244, $L_N(\theta^\circ) = \frac{\lambda^2}{2}\big\|(K^\circ + \lambda)^{-1}y\big\|^2_N$.*

Proof. ☐

Theorem 249. *Along a solution of the gradient flow (**??**), $t \mapsto J_N(\theta_t)$ is nonincreasing on $[0, \infty)$, since $\frac{\mathrm{d}}{\mathrm{d}t}J_N(\theta_t) = -M\|\nabla J_N(\theta_t)\|^2 \le 0$.*

Proof. ☐

Theorem 250. *Along a solution of the gradient flow (**??**), for all $t \ge 0$, $\frac{\lambda}{2M}\sum_\ell a_\ell(t)^2 + \frac{1}{M}\sum_\ell V(z_\ell(t)) \le J_N(\theta_0)$. (With $V \ge 0$ and (A6) this bounds the trajectory in a compact set.)*

Proof. ☐

Theorem 251. *Assume (A3), (A6) and (A7): $z \mapsto \varphi_z(x)$ and V are real analytic, $V \ge 0$ is coercive and Z is finite dimensional. For every $\theta_0 \in \Theta^M$ the gradient flow (**??**) has a unique solution on $[0, \infty)$, which is bounded, has finite length and converges to a stationary point θ°: $\int_0^\infty \|\dot\theta_t\|\,\mathrm{d}t < \infty$, $\lim_{t\to\infty}\theta_t = \theta^\circ$, $\nabla J_N(\theta^\circ) = 0$.*

Proof. ☐

4.3 Without amplitude regularization the null-space component survives

Theorem 252. $S_N S_N^\dagger y = P_{\operatorname{ran} S_N} y$.

Proof. □

Theorem 253. *For $\lambda \geq 0$ and a solution γ_t of (**??**), $P_{\ker S_N}\gamma_t = e^{-\lambda t} P_{\ker S_N}\gamma_0$ for all t.*

Proof. □

Theorem 254. *For $\lambda = 0$ and a solution γ_t of (**??**), $P_{\ker S_N}\gamma_t = P_{\ker S_N}\gamma_0$ for all $t \geq 0$: the null-space component is conserved by the flow.*

Proof. □

Theorem 255. *For $\lambda > 0$ and a solution γ_t of (**??**), $\gamma_t = \gamma_\lambda + e^{-t(B+\lambda)}(\gamma_0 - \gamma_\lambda)$ with $\gamma_\lambda = (B + \lambda)^{-1} S_N^* y \in (\ker S_N)^\perp$; the endpoint does not depend on γ_0.*

Proof. □

Theorem 256. *For $\lambda > 0$ and a solution γ_t of (**??**), $\|\gamma_t - \gamma_\lambda\|_M \leq e^{-\lambda t}\|\gamma_0 - \gamma_\lambda\|_M$ for $t \geq 0$.*

Proof. □

Theorem 257. *$B = S_N^* S_N$ is invertible on $(\ker S_N)^\perp$: there is $\sigma > 0$ (the smallest positive eigenvalue $\sigma_{\min}$ of B is the largest such) with $\sigma\|u\|^2 \leq \langle u, Bu\rangle$ for all $u \in (\ker S_N)^\perp$.*

Proof. □

Theorem 258. *For $\lambda = 0$ and a solution γ_t of (**??**), $\gamma_t = S_N^\dagger y + P_{\ker S_N}\gamma_0 + e^{-tB}(P_{(\ker S_N)^\perp}\gamma_0 - S_N^\dagger y)$ (**??**).*

Proof. □

Theorem 259. *For $\lambda = 0$, a solution γ_t of (**??**), $\gamma_\infty := S_N^\dagger y + P_{\ker S_N}\gamma_0$ and any $\sigma > 0$ with $\sigma\|u\|^2 \leq \langle u, Bu\rangle$ on $(\ker S_N)^\perp$ (in particular $\sigma = \sigma_{\min}$), $\|\gamma_t - \gamma_\infty\|_M \leq e^{-\sigma t}\|P_{(\ker S_N)^\perp}\gamma_0 - S_N^\dagger y\|_M$ for $t \geq 0$.*

Proof. □

Theorem 260. *For $\lambda = 0$ and a solution γ_t of (**??**), $\gamma_\infty := \lim_{t\to\infty}\gamma_t = S_N^\dagger y + P_{\ker S_N}\gamma_0$.*

Proof. □

Theorem 261. *For $\gamma_\infty = S_N^\dagger y + P_{\ker S_N}\gamma_0$, $S_N\gamma_\infty = P_{\operatorname{ran} S_N} y$.*

Proof. □

Theorem 262. *For $\gamma_\infty = S_N^\dagger y + P_{\ker S_N}\gamma_0$, $\|\gamma_\infty - S_N^\dagger y\|_M = \|P_{\ker S_N}\gamma_0\|_M$, and $\gamma_\infty = S_N^\dagger y$ if and only if $P_{\ker S_N}\gamma_0 = 0$.*

Proof. □

Theorem 263. *For $\sigma > 0$ with $\sigma\|u\|^2 \leq \langle u, Bu\rangle$ on $(\ker S_N)^\perp$ and $\lambda > 0$, $\|\gamma_\lambda - S_N^\dagger y\|_M \leq \lambda\|S_N^\dagger y\|_M/\sigma$; in particular $\gamma_\lambda \to S_N^\dagger y$ as $\lambda \downarrow 0$.*

Proof. ☐

Theorem 264. *For $\lambda > 0$, $\gamma_\lambda = (B+\lambda)^{-1}S_N^* y = S_N^*(K_{N,\nu_M}+\lambda)^{-1}y$, the values at the particle positions of $R^{(N)}_{\lambda,\nu_M} y$.*

Proof. ☐

Theorem 265. $\operatorname{rank} S_N \le N$*, hence* $\dim\ker S_N \ge M-N$*; in particular* $\dim\ker S_N \ge 1$ *if* $M > N$.

Proof. ☐

Theorem 266. *If the components of γ_0 are centered, pairwise uncorrelated with variance τ^2 (e.g. i.i.d.) and independent of $z_1, \dots, z_M$, then $\mathbb{E}[\|P_{\ker S_N}\gamma_0\|_M^2 \mid z] = \frac{\tau^2 \dim\ker S_N}{M} \ge \tau^2(1-\frac{N}{M})$* **(??)**.

Proof. ☐

Theorem 267. *Under the assumptions of 266, for every fixed test vector $q \in \mathcal{H}_M$, $\mathbb{E}[\langle q, P_{\ker S_N}\gamma_0\rangle_M^2 \mid z] = \frac{\tau^2}{M}\|P_{\ker S_N} q\|_M^2 \le \frac{\tau^2\|q\|_M^2}{M}$* **(??)**.

Proof. ☐

Theorem 268. *(Necessity of amplitude regularization.) For $\lambda = 0$, $\beta = 0$ and frozen hidden parameters, the endpoint of the fixed-feature flow is $\gamma_\infty = S_N^\dagger y + P_{\ker S_N}\gamma_0$: the canonical empirical ridgelet transform plus the* $\ker S_N$ *component of the initialization, which is data independent, invisible to the training loss and conserved by the flow. With amplitude regularization $\lambda > 0$ the* $\ker S_N$ *component decays like $e^{-\lambda t}$: $P_{\ker S_N}\gamma_t = e^{-\lambda t}P_{\ker S_N}\gamma_0$.*

Proof. ☐

Chapter 5

Sample-size threshold: from the empirical to the population Gibbs minimizer

Section 5 of the paper. The empirical problem is the population problem for the empirical feature on the index space of the sample, so the structure theorem applies verbatim (`lem:empirical-gibbs`); uniform deviations over the class of measures with bounded first amplitude moment are controlled by the Rademacher complexity of the feature class (`lem:uniform-deviation`), which gives the sample-size threshold (`thm:m3`).

5.1 Empirical Gibbs structure, moment bounds and tails

Theorem 269. *For the empirical feature, $F_\rho(i) = F_\rho(x_i)$: the realization is the vector $F_\rho|_x = (F_\rho(x_i))_i \in \mathbb{R}^N$.*

Proof. $\square$

Theorem 270. *$L_N(\rho) = \frac{1}{2N}\sum_i (F_\rho(x_i) - y_i)^2 = \frac{1}{2}\|F_\rho|_x - y\|_N^2$ is the population risk of the empirical feature for the uniform measure on $\{1, \dots, N\}$ and the target y.*

Proof. $\square$

Theorem 271. *$\mathcal{F}_N(\rho) = L_N(\rho) + \beta\,\mathrm{KL}(\rho\|\mu_U)$ is the free energy of the empirical feature for the uniform measure on $\{1, \dots, N\}$ and the target y.*

Proof. $\square$

Theorem 272. *For $r \in \mathbb{R}^N$ the analysis map of the empirical feature is $(S^*r)(z) = \frac{1}{N}\sum_i \varphi_z(x_i) r_i = (S_N^* r)(z)$.*

Proof. $\square$

Theorem 273. *For $\nu \in \mathcal{P}(Z)$ and $h \in \mathbb{R}^N$ the kernel operator of the empirical feature is $K_{N,\nu}$: $(K_\nu h)_i = \int \varphi_z(x_i)\langle \Phi_z, h\rangle_N \, \nu(\mathrm{d}z)$.*

Proof. $\square$

Theorem 274. *For $\lambda > 0$ and $\nu \in \mathcal{P}(Z)$, $K_{N,\nu} + \lambda$ is injective (hence bijective) on $\mathbb{R}^N$, since $K_{N,\nu}$ is positive semidefinite for $\langle\cdot,\cdot\rangle_N$.*

Proof. ☐

Theorem 275. *Let s be measurable with $|s| \le C$, $\rho_s = \hat{\mu}_{W_s}$ as in Lemma 218, $B_C := \frac{C}{\lambda} + \sqrt{\frac{2\beta}{\pi\lambda}}$ and $T \ge C/\lambda$. Then $\int_{\{|a|>T\}} |a| \,\mathrm{d}\rho_s \le B_C \exp(-\frac{\lambda}{2\beta}(T - \frac{C}{\lambda})^2)$. (Conditionally on z, $a \sim \mathcal{N}(-s(z)/\lambda, \beta/\lambda)$ with $|s(z)/\lambda| \le C/\lambda$; apply Lemma 908 with $m_0 = C/\lambda$, $v = \beta/\lambda$ and integrate over ν_{ρ_s}.)*

Proof. Step 1: the set integral as the integral of ‘g(a) = |a| 1_|a| > T‘.
Step 2: transport through the disintegration ‘ρ_s.map swap = ν ⊗ κ_s‘.
Step 3: the inner Gaussian truncated moment, uniformly in ‘z‘.
Step 4: integrate over the hidden marginal. ☐

Theorem 276. *The minimizer ρ_N^* of $\mathcal{F}_N$ on $\mathcal{D}$ is the minimizer of the free energy of the empirical feature $(z, i) \mapsto \varphi_z(x_i)$ for the uniform measure on $\{1, \dots, N\}$ and the target $y \in \mathbb{R}^N = L^2$(uniform); all of Lemma* **??** *and Theorem* **??** *apply to it.*

Proof. ☐

Theorem 277. *With $r_N^* := F_{\rho_N^*}|_x - y$ and $g_N := S_N^* r_N^*$, $\rho_N^* = \hat{\mu}_{W_N^*}$ relative to μ_U with $W_N^*(\theta) = a\, g_N(z)$, i.e. $\rho_N^*(\,\mathrm{d}\theta) \propto e^{-a g_N(z)/\beta} \mu_U(\,\mathrm{d}\theta)$.*

Proof. ☐

Theorem 278. *Let $r_N^* := F_{\rho_N^*}|_x - y \in \mathbb{R}^N$, $g_N := S_N^* r_N^*$ and $\nu_N^* := \nu_{\rho_N^*}$. Then $\rho_N^*(\,\mathrm{d}a\,\mathrm{d}z) = \nu_N^*(\,\mathrm{d}z)\,\mathcal{N}(-g_N(z)/\lambda, \beta/\lambda)(\,\mathrm{d}a)$.*

Proof. ☐

Theorem 279. *$\nu_N^*(\,\mathrm{d}z) \propto \exp(\frac{g_N(z)^2}{2\lambda\beta})\nu_0(\,\mathrm{d}z)$ (with $\nu_0 \propto e^{-V/\beta}\,\mathrm{d}z$ this is $\nu_N^*(\,\mathrm{d}z) \propto \exp(\frac{g_N(z)^2}{2\lambda\beta} - \frac{V(z)}{\beta})\,\mathrm{d}z$).*

Proof. ☐

Theorem 280. *$m_{\rho_N^*}(z) = -g_N(z)/\lambda$ for ν_N^*-a.e. z.*

Proof. ☐

Theorem 281. *With $K := K_{N,\nu_N^*}$,*

$$(K+\lambda)r_N^* = -\lambda y, \qquad F_{\rho_N^*}|_x = -\lambda^{-1} K r_N^*, \qquad m_{\rho_N^*} = R^N_{\lambda,\nu_N^*} y \ (\nu_N^*\text{-a.e.}), \qquad g_N = -\lambda R^N_{\lambda,\nu_N^*} y,$$

and $K + \lambda$ is injective on $\mathbb{R}^N$, so that $r_N^* = -\lambda(K+\lambda)^{-1} y$ and $F_{\rho_N^*}|_x = K(K+\lambda)^{-1} y$.

Proof. Step 1: ‘(K_ν + λ) r = −λ y‘ in ‘L²(uniformFin N)‘, from ‘r = −λ (K_ν + λ)⁻¹ y‘.
Step 2: transport to ‘ℝ^N‘ through the dictionary.
Step 3: ‘F_ρ|_x = r + y = −λ⁻¹ K_ν r‘. ☐

Theorem 282. *If $|y_i| \le Y_{\max}$ for all i, then deterministically $\|r_N^*\|_N \le \|y\|_N \le Y_{\max}$, $\sup_z |g_N(z)| \le Y_{\max}$ and $|m_{\rho_N^*}(z)| \le Y_{\max}/\lambda$ for ν_N^*-a.e. z.*

Proof. ☐

Theorem 283. *If $|y_i| \le Y_{\max}$ for all i, then deterministically $\int |a| \, d\rho_N^* \le B := \frac{Y_{\max}}{\lambda} + \sqrt{\frac{2\beta}{\pi\lambda}}$ and $\int a^2 \, d\rho_N^* \le \frac{Y_{\max}}{\lambda^2}^2 + \frac{\beta}{\lambda}$.*

Proof. □

Theorem 284. *If $|y_i| \le Y_{\max}$ for all i and $T \ge Y_{\max}/\lambda$, then deterministically $\int_{\{|a|>T\}} |a| \, d\rho_N^* \le B \exp(-\frac{\lambda}{2\beta}(T - \frac{Y_{\max}}{\lambda})^2)$, $B = \frac{Y_{\max}}{\lambda} + \sqrt{\frac{2\beta}{\pi\lambda}}$.*

Proof. □

Theorem 285. *If $\|f\|_{L^2(P_X)} \le Y_{\max}$ then, with $g^* := S^* r^*$ as in Theorem **??**, $\|r^*\|_{L^2(P_X)} \le \|f\|_{L^2(P_X)} \le Y_{\max}$, $\sup_z |g^*(z)| \le Y_{\max}$ and $|m_{\rho^*}(z)| \le Y_{\max}/\lambda$ for ν^*-a.e. z.*

Proof. □

Theorem 286. *If $\|f\|_{L^2(P_X)} \le Y_{\max}$ then $\int |a| \, d\rho^* \le B = \frac{Y_{\max}}{\lambda} + \sqrt{\frac{2\beta}{\pi\lambda}}$ and $\int a^2 \, d\rho^* \le \frac{Y_{\max}}{\lambda^2}^2 + \frac{\beta}{\lambda}$.*

Proof. □

Theorem 287. *If $\|f\|_{L^2(P_X)} \le Y_{\max}$ and $T \ge Y_{\max}/\lambda$, then $\int_{\{|a|>T\}} |a| \, d\rho^* \le B \exp(-\frac{\lambda}{2\beta}(T - \frac{Y_{\max}}{\lambda})^2)$.*

Proof. □

Theorem 288. *Let f be a regression function of Q with $|f| \le C$, $|Y| \le Y_{\max}$ a.s., and ρ with $\int |a| \, d\rho < \infty$. Then $\tilde{L}(\rho) = L(\rho) + \frac{1}{2}\mathbb{E}(Y - f(X))^2 = L(\rho) + \frac{1}{2}\mathbb{E}\operatorname{Var}(Y \mid X)$, since $\mathbb{E}[(F_\rho(X) - f(X))(f(X) - Y)] = 0$ by the tower property.*

Proof. Step 1: pointwise, '(F_ρ − Y)² = g² − 2 (Y − f) g + (Y − f)²' with 'g = F_ρ − f'.
Step 2: integrability of the three terms.
Step 3: the tower property kills the cross term, and '∫ g² dQ = ∫ (F_ρ − f)² dP_X'. □

Theorem 289. *Under the hypotheses of Lemma 288, for ρ, ρ' with integrable amplitude, $\tilde{L}(\rho) - \tilde{L}(\rho') = L(\rho) - L(\rho')$.*

Proof. □

Theorem 290. *Let f be a regression function of Q with $|f| \le C$, $|Y| \le Y_{\max}$ a.s., P_X the first marginal of Q, and ρ^* the minimizer of $\mathcal{F}$ on $\mathcal{D}$ (for the target $f \in L^2(P_X)$). Then for every $\rho \in \mathcal{D}$, $\mathrm{KL}(\rho\|\rho^*) < \infty$ and*

$$\tilde{\mathcal{F}}(\rho) - \tilde{\mathcal{F}}(\rho^*) = \tfrac{1}{2}\|F_\rho - F_{\rho^*}\|^2_{L^2(P_X)} + \beta \, \mathrm{KL}(\rho\|\rho^*).$$

(Lemma 212 and Lemma 288: $\tilde{\mathcal{F}} - \mathcal{F}$ is the constant $\frac{1}{2}\mathbb{E}(Y - f(X))^2$.)

Proof. The population risk only sees the a.e. class of the target. □

Theorem 291. *Let ρ_N^* be the minimizer of $\mathcal{F}_N$ on $\mathcal{D}$ for the sample $(x_i, y_i)_{i=1}^N$. Then for every $\rho \in \mathcal{D}$, $\mathrm{KL}(\rho\|\rho_N^*) < \infty$ and*

$$\mathcal{F}_N(\rho) - \mathcal{F}_N(\rho_N^*) = \tfrac{1}{2}\|F_\rho|_x - F_{\rho_N^*}|_x\|_N^2 + \beta \, \mathrm{KL}(\rho\|\rho_N^*).$$

(Lemma 212 for the empirical feature, through the dictionary of Lemma 61.)

Proof. □

Theorem 292. *For a sample $\omega = (x_i, y_i)_{i=1}^N$ and the minimizer ρ_N^* of $\mathcal{F}_N$ on $\mathcal{D}$, every $\rho \in \mathcal{D}$ satisfies $\mathcal{F}_N(\rho) - \mathcal{F}_N(\rho_N^*) = \frac{1}{2}\|F_\rho|_x - F_{\rho_N^*}|_x\|_N^2 + \beta \, \mathrm{KL}(\rho\|\rho_N^*)$.*

Proof. □

5.2 Uniform deviation over the class of bounded first amplitude moment

Theorem 293. *For $\rho \in \mathcal{P}_B$ and every $x \in \mathcal{X}$, $|F_\rho(x)| \leq B$.*

Proof. □

Theorem 294. *For $|a| \leq B$ and $z \in Z$, $\delta_{(a,z)} \in \mathcal{P}_B$.*

Proof. □

Theorem 295. *Let $B \geq 0$, $x_1, \dots, x_N \in \mathcal{X}$ and $\sigma \in \{\pm 1\}^N$, and let $\psi_\sigma(z) := \frac{1}{N}\sum_i \sigma_i \varphi_z(x_i)$. Then $\sup_{\rho \in \mathcal{P}_B} \frac{1}{N}\sum_i \sigma_i F_\rho(x_i) = B \sup_{z \in Z} |\psi_\sigma(z)|$ (and the same with $\left|\frac{1}{N}\sum_i \sigma_i F_\rho(x_i)\right|$ on the left).*

Proof. Upper bound: '|(1/N) ∑ σ_k F_ρ(x_k)| = |∫ a ψ_σ(z) dρ| ≤ C ∫ |a| dρ ≤ B C'.
Lower bound: 'ρ = δ_(B s, z)' with 's = sign ψ_σ(z)' gives the value 'B |ψ_σ(z)|'. □

Theorem 296. *Let $B \geq 0$ and $x_1, \dots, x_N \in \mathcal{X}$. Then, in the one-sided convention of the paper, $\widehat{\mathfrak{R}}_N(\mathcal{H}_B) = B\widehat{\mathfrak{R}}_N(\Phi)$ for $\mathcal{H}_B = \{F_\rho : \rho \in \mathcal{P}_B\}$, where $\widehat{\mathfrak{R}}_N(\Phi) = \mathbb{E}_\sigma \sup_z |\psi_\sigma(z)|$ is 'featureRademacher'.*

Proof. □

Theorem 297. *Under the hypotheses of Lemma 296, also in the absolute convention $\widehat{\mathfrak{R}}_N(\mathcal{H}_B) = B\widehat{\mathfrak{R}}_N(\Phi)$.*

Proof. □

Theorem 298. *If for every $z \in Z$ there is $z' \in Z$ with $\varphi_{z'} = -\varphi_z$ (for $\tanh$, $z' = -z$), then $\widehat{\mathfrak{R}}_N(\Phi)$ coincides with the one-sided complexity $\mathbb{E}_\sigma \sup_z \frac{1}{N}\sum_i \sigma_i \varphi_z(x_i)$ of the paper.*

Proof. □

Theorem 299. *$\mathcal{P}_B^0$ is countable.*

Proof. □

Theorem 300. *Let Z be first countable, $Z_0 \subseteq Z$ dense, φ continuous in z, $Q \in \mathcal{P}(\mathcal{X} \times \mathbb{R})$, $B \geq 0$ and $x_1, \dots, x_N \in \mathcal{X}$. For every $\rho \in \mathcal{P}_B$ and $\varepsilon > 0$ there is $\rho' \in \mathcal{P}_B^0$ with $\int |F_\rho - F_{\rho'}| \, dQ \leq \varepsilon$ and $|F_\rho(x_k) - F_{\rho'}(x_k)| \leq \varepsilon$ for all k. The proof samples M particles from ρ reweighted by $|a|$ (Monte Carlo, expected squared error $\leq B^2/M$), then moves the atoms into Z_0 and the amplitude into $\mathbb{Q}$.*

Proof. 'B' = 0': 'F_ρ = 0', approximated by 'δ_(0, z$_0$)'.
'B' > 0': Monte Carlo approximation, then atoms to 'Z$_0$' and amplitude to 'ℚ'. □

Theorem 301. *Let Z be separable and first countable, φ continuous in z, $Q \in \mathcal{P}(\mathcal{X} \times \mathbb{R})$ with $|Y| \leq Y_{\max}$ a.s., $B \geq 0$, $Y_{\max} \geq 0$ and $N \geq 1$. Then $\mathbb{E}\, G_\pm \leq 2B(B + Y_{\max})\, \mathfrak{R}_N(\Phi)$.*

Proof. □

Theorem 302. *Let Z be separable and first countable, φ continuous in z, $Q \in \mathcal{P}(\mathcal{X} \times \mathbb{R})$ with $|Y| \leq Y_{\max}$ a.s., $B > 0$, $Y_{\max} \geq 0$ and $N \geq 1$. Let $G := \sup_{\rho \in \mathcal{P}_B} |L_N(\rho) - \tilde{L}(\rho)|$. Then for every $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the i.i.d. sample of size N from Q,*

$$G \leq 2B(B + Y_{\max})\, \mathfrak{R}_N(\Phi) + \frac{(B + Y_{\max})^2}{2} \sqrt{\frac{\log(2/\delta)}{2N}}.$$

The proof is symmetrization (one-sided), the contraction principle, the identity $\widehat{\mathfrak{R}}_N(\mathcal{H}_B) = B\widehat{\mathfrak{R}}_N(\Phi)$, McDiarmid's inequality for $G_\pm$ with bounded differences $(B + Y_{\max})^2/(2N)$ and a union bound; the suprema over $\mathcal{P}_B$ are computed on the countable dense subclass $\mathcal{P}_B^0$.

Proof. Step 4 (concentration): McDiarmid for a measurable function with bounded differences 'b/N', at the level 'ε', gives failure probability 'δ/2'.

The union bound and the complement.

The good event: all labels in '[−Ymax, Ymax]', where 'G = G̃'. □

Theorem 303. *Assume (A3) and an i.i.d. sample of size $N \geq 1$ (only the x_i are used). For every $\delta_2 \in (0, 1)$, with probability at least $1 - \delta_2$, $H_N \leq 2\,\mathfrak{R}_N(\Psi) + \sqrt{2\log(2/\delta_2)/N}$. (FoML's bound has the sharper radius $\sqrt{2\log(1/\delta_2)/N}$, obtained without a union bound.)*

Proof. □

Theorem 304. *For $|e| \leq E_*$ and every sample $(x_i, y_i)_{i=1}^N$, $\widehat{\mathfrak{R}}_N(\{\varphi_z e\}) \leq E_*\, \widehat{\mathfrak{R}}_N(\Phi)$, by the contraction lemma with the E_*-Lipschitz maps $t \mapsto e(x_i, y_i)t$.*

Proof. □

Theorem 305. *Assume (A3), an i.i.d. sample of size $N \geq 1$ and a measurable $e\colon \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ with $|e| \leq E_*$, $E_* > 0$ (in the paper $e(x, y) = y - f(x)$). For every $\delta_1 \in (0, 1)$, with probability at least $1 - \delta_1$, $G_N \leq 2E_*\,\mathfrak{R}_N(\Phi) + E_*\sqrt{2\log(2/\delta_1)/N}$.*

Proof. □

Theorem 306. *Under the hypotheses of Lemma 305, with $e(x, y) = y - f(x)$ for a measurable f and $|Y - f(X)| \leq E_*$ almost surely, the same bound holds for $G_N = \sup_z |(P_N - P)(\varphi_z e)|$.*

Proof. On the samples with '|e(x_k, y_k)| ≤ E_*' the two deviations agree. □

5.3 Complexity of the class of tanh ridge functions

Theorem 307. *For every m, $\mathrm{HC}(m, 8, 6)$ holds: for every N, every $x_1, \ldots, x_N \in \mathbb{R}^m$ and every $0 < \varepsilon \leq 1/2$, $\mathcal{N}(\varepsilon, \Phi, L_2(P_N)) \leq (8/\varepsilon)^{6(m+1)}$.*

Proof. Lemma 825 with $d = m + 1$ (Lemma 110) and $|\varphi_z| \leq 1$. □

Theorem 308. *For every m, N and every $x_1, \ldots, x_N \in \mathbb{R}^m$, $\widehat{\mathfrak{R}}_N(\Phi) \leq 18\sqrt{(m + 1)\log(N + 1)/N}$.*

Proof. Lemma 815 with $d = m + 1$ (Lemma 110); for $N = 0$ both sides vanish. Compared with Lemma 314(b), the factor $\sqrt{\log(N + 1)}$ comes from the factor N in the Sauer–Shelah covering bound, which Haussler's bound removes. □

Theorem 309. *Assume (A3). For every N and every $x_1, \ldots, x_N$, $\widehat{\mathfrak{R}}_N(\Psi) \leq 26\sqrt{(m + 1)\log(N + 1)/N}$.*

Proof. Lemma 816 with $F = G = \Phi$ and $d = m + 1$ (Lemma 110); for $N = 0$ both sides vanish. $\square$

Theorem 310. *Assume* $\mathrm{HC}(m, A, c)$ *for all* m*, with* $2A \geq 1$*,* $c \geq 1$*. Then for every* m*,* N *and* $x_1, \dots, x_N \in \mathbb{R}^m$*,* $\widehat{\mathfrak{R}}_N(\Phi) \leq (1 + 6\sqrt{\log(2A)} + 12\sqrt{2})\sqrt{c}\,\sqrt{(m+1)/N}$*, i.e. Lemma 314(b) with* $C_{\mathrm{P}} = (1 + 6\sqrt{\log(2A)} + 12\sqrt{2})\sqrt{c}$.

Proof. Lemma 818 with $d = c(m + 1)$; for $N = 0$ both sides vanish. $\square$

Theorem 311. *Assume* $\mathrm{HC}(m, A, c)$ *for all* m*, with* $2A \geq 1$*,* $c \geq 1$*. Then for every* m*,* N *and* $x_1, \dots, x_N \in \mathbb{R}^m$*,* $\widehat{\mathfrak{R}}_N(\Psi) \leq (1 + 6\sqrt{\log(4A)} + 12\sqrt{2})\sqrt{2c}\,\sqrt{(m+1)/N}$.

Proof. Lemma 813 gives $\mathcal{N}(\varepsilon, \Psi, L_2(P_N)) \leq \mathcal{N}(\varepsilon/2, \Phi, L_2(P_N))^2 \leq (2A/\varepsilon)^{2c(m+1)}$, and Lemma 818 applies with $d = 2c(m + 1)$. $\square$

Theorem 312. *For every* m*,* N *and* $x_1, \dots, x_N \in \mathbb{R}^m$*,* $\widehat{\mathfrak{R}}_N(\Phi) \leq (1+6\sqrt{\log 16}+12\sqrt{2})\sqrt{6}\,\sqrt{(m+1)/N}$*; numerically the constant is at most* 69.

Proof. Lemma 310 with Lemma 307. $\square$

Theorem 313. *For every* m*,* N *and* $x_1, \dots, x_N \in \mathbb{R}^m$*,* $\widehat{\mathfrak{R}}_N(\Psi) \leq (1+6\sqrt{\log 32}+12\sqrt{2})\sqrt{12}\,\sqrt{(m+1)/N}$*; numerically the constant is at most* 101.

Proof. Lemma 311 with Lemma 307. $\square$

Theorem 314. *(a) The pseudo-dimension of the class* $\Phi = \{\varphi_z : z \in Z\}$ *of* tanh *ridge functions* $x \mapsto \tanh(w^\top x - b)$*,* $(w, b) \in \mathbb{R}^m \times \mathbb{R}$*, is at most* $m + 1$*. (b) There is an absolute constant* C_{P} *such that for every* N *and every* $x_1, \dots, x_N$*,* $\widehat{\mathfrak{R}}_N(\Phi) \leq C_{\mathrm{P}}\sqrt{(m+1)/N}$*, hence* $\mathfrak{R}_N(\Phi) \leq C_{\mathrm{P}}\sqrt{(m+1)/N}$*. A crude evaluation through Haussler's covering bound and Dudley's entropy integral gives* $C_{\mathrm{P}} \leq 43$.

Proof. $\square$

Theorem 315. *Assume (A3). For every* N *and every* $x_1, \dots, x_N$*,* $\widehat{\mathfrak{R}}_N(\Psi) \leq C'_{\mathrm{P}}\sqrt{(m+1)/N}$ *with* $C'_{\mathrm{P}} := \min\{4C_{\mathrm{P}}, 63\}$.

Proof. $\square$

Theorem 316. *Under Lemma 314, for every* $Q \in \mathcal{P}(\mathcal{X} \times \mathbb{R})$*,* $\mathfrak{R}_N(\Phi) \leq C_{\mathrm{P}}\sqrt{(m+1)/N}$.

Proof. $\square$

Theorem 317. *Under Lemma 315, for every* $Q \in \mathcal{P}(\mathcal{X} \times \mathbb{R})$*,* $\mathfrak{R}_N(\Psi) \leq C'_{\mathrm{P}}\sqrt{(m+1)/N}$.

Proof. $\square$

Theorem 318. *For every* $Q \in \mathcal{P}(\mathcal{X} \times \mathbb{R})$*,* $\mathfrak{R}_N(\Phi) \leq 18\sqrt{(m+1)\log(N+1)/N}$.

Proof. $\square$

Theorem 319. *For every* $Q \in \mathcal{P}(\mathcal{X} \times \mathbb{R})$*,* $\mathfrak{R}_N(\Psi) \leq 26\sqrt{(m+1)\log(N+1)/N}$.

Proof. $\square$

5.4 The main estimate, the sample-size threshold and the convergence of the learned quantities

Theorem 320. *$\rho^* \in \mathcal{P}_B$ (Lemma **??**(c)).*

Proof. □

Theorem 321. *If $|y_i| \le Y_{\max}$ for all i then $\rho_N^*(\omega) \in \mathcal{P}_B$ (Lemma **??**(c)).*

Proof. □

Theorem 322. *Under the hypotheses of Lemma 302, for every $\delta \in (0,1)$, $\mathbb{P}(E_\delta) \ge 1 - \delta$.*

Proof. □

Theorem 323. *On a sample with $|y_i| \le Y_{\max}$, $\tilde{\mathcal{F}}(\rho_N^*) - \tilde{\mathcal{F}}(\rho^*) = [\tilde{L}(\rho_N^*) - L_N(\rho_N^*)] + [\mathcal{F}_N(\rho_N^*) - \mathcal{F}_N(\rho^*)] + [L_N(\rho^*) - \tilde{L}(\rho^*)] \le G_+ + G_-$, since the middle term is ≤ 0 by minimality of ρ_N^* and $\rho_N^*, \rho^* \in \mathcal{P}_B$.*

Proof. □

Theorem 324. *On a sample with $|y_i| \le Y_{\max}$, $\mathcal{F}_N(\rho^*) - \mathcal{F}_N(\rho_N^*) = [L_N(\rho^*) - \tilde{L}(\rho^*)] + [\tilde{\mathcal{F}}(\rho^*) - \tilde{\mathcal{F}}(\rho_N^*)] + [\tilde{L}(\rho_N^*) - L_N(\rho_N^*)] \le G_+ + G_-$, since the middle term is ≤ 0 by minimality of ρ^*.*

Proof. □

Theorem 325. *On the event E_δ of Lemma 302,*

$$\tfrac{1}{2}\|F_{\rho_N^*} - F_{\rho^*}\|^2_{L^2(P_X)} + \beta\,\mathrm{KL}(\rho_N^*\|\rho^*) = \tilde{\mathcal{F}}(\rho_N^*) - \tilde{\mathcal{F}}(\rho^*) \le G_+ + G_- \le 2G \le \Delta_N(\delta).$$

Proof. □

Theorem 326. *On the event E_δ,*

$$\tfrac{1}{2}\|F_{\rho_N^*}|_x - F_{\rho^*}|_x\|^2_N + \beta\,\mathrm{KL}(\rho^*\|\rho_N^*) = \mathcal{F}_N(\rho^*) - \mathcal{F}_N(\rho_N^*) \le G_+ + G_- \le 2G \le \Delta_N(\delta).$$

Proof. □

Theorem 327. *On the event E_δ, $\beta\,\mathrm{KL}(\rho_N^*\|\rho^*) \le \Delta_N(\delta)$ and $\beta\,\mathrm{KL}(\rho^*\|\rho_N^*) \le \Delta_N(\delta)$.*

Proof. □

Theorem 328. *Assume the setting of Section **??** ($|Y| \le Y_{\max}$, i.i.d. sample, $\lambda, \beta > 0$, Z separable and first countable, φ continuous in z) and Lemma **??** for L and L_N. Let B be as in (82) and $\Delta_N(\delta)$ as in (83). For every $\delta \in (0,1)$, with probability at least $1-\delta$ over the sample, the following hold simultaneously:*

$$\tfrac{1}{2}\|F_{\rho_N^*} - F_{\rho^*}\|^2_{L^2(P_X)} + \beta\,\mathrm{KL}(\rho_N^*\|\rho^*) \le \Delta_N(\delta), \qquad \tfrac{1}{2}\|F_{\rho_N^*}|_x - F_{\rho^*}|_x\|^2_N + \beta\,\mathrm{KL}(\rho^*\|\rho_N^*) \le \Delta_N(\delta).$$

Proof. □

Theorem 329. *If $\mathfrak{R}_N(\Phi) \le C_{\mathrm{P}}\sqrt{(m+1)/N}$ (Lemma 314) then $\Delta_N(\delta) \le D(\delta)/\sqrt{N}$ with $D(\delta)$ as in (84).*

Proof. □

Theorem 330. *In the setting of Theorem 328, let $\varepsilon > 0$, $\delta \in (0,1)$, $\mathfrak{R}_N(\Phi) \le C_{\mathrm{P}}\sqrt{(m+1)/N}$ and $N_0(\varepsilon,\delta)$ as in (85). If $N \ge N_0(\varepsilon,\delta)$ then on E_δ (probability at least $1-\delta$), $\mathrm{KL}(\rho_N^*\|\rho^*) \le \varepsilon$ and $\mathrm{KL}(\rho^*\|\rho_N^*) \le \varepsilon$.*

Proof. □

Theorem 331. *On E_δ, $\tilde{\mathcal{F}}(\rho_N^*) - \tilde{\mathcal{F}}(\rho^*) \le \Delta_N(\delta)$ and $\mathcal{F}_N(\rho^*) - \mathcal{F}_N(\rho_N^*) \le \Delta_N(\delta)$.*

Proof. □

Theorem 332. *On E_δ, $\|F_{\rho_N^*} - F_{\rho^*}\|_{L^2(P_X)} \le \sqrt{2\Delta_N(\delta)}$ and $\|F_{\rho_N^*}|_x - F_{\rho^*}|_x\|_N \le \sqrt{2\Delta_N(\delta)}$.*

Proof. □

Theorem 333. *On E_δ, $|\tilde{L}(\rho_N^*) - \tilde{L}(\rho^*)| = |L(\rho_N^*) - L(\rho^*)| \le \Delta_N(\delta) + Y_{\max}\sqrt{2\Delta_N(\delta)}$.*

Proof. Step 1: the expansion ‘L(ρ_N^*) − L(ρ*) = ⟪F_ρ_N^* − F_ρ*, r*⟫ + ½‖F_ρ_N^* − F_ρ*‖²‘.
Step 2: ‘‖r*‖ ≤ Ymax‘, ‘‖d‖ ≤ √(2Δ)‘ and ‘½‖d‖² ≤ Δ‘. □

Theorem 334. *$(S^*r^*)(z) = \mathbb{E}[\varphi_z(X)(F_{\rho^*}(X) - f(X))] = \mathbb{E}[\varphi_z(X)(F_{\rho^*}(X) - Y)]$ by the tower property.*

Proof. Step 1: ‘S* r* (z) = ∫ φ_z(x) (F_ρ*(x) − f(x)) P_X(dx) = ∫ φ_z(x) (F_ρ*(x) − f(x)) Q(dx dy)‘.
Step 2: split ‘F_ρ* − f = (F_ρ* − y) + (y − f)‘; the second term integrates to ‘0‘. □

Theorem 335. *On $E_{\delta/2}$ (in particular on E'_δ), with $\Delta_N := \Delta_N(\delta/2)$, $\|\rho_N^* - \rho^*\|_{\mathrm{TV}} \le \sqrt{\Delta_N/(2\beta)}$ and $\|\nu_N^* - \nu^*\|_{\mathrm{TV}} \le \sqrt{\Delta_N/(2\beta)}$ (Pinsker's inequality and (**??**); the total variation of the marginals is at most that of the joint laws).*

Proof. □

Theorem 336. *On a sample with $|y_i| \le Y_{\max}$, for every $z \in Z$,*

$$g_N(z) - g^*(z) = \frac{1}{N}\sum_i \varphi_z(x_i)(F_{\rho_N^*}(x_i) - F_{\rho^*}(x_i)) - \Big[\frac{1}{N}\sum_i \varphi_z(x_i)e(x_i,y_i) - \mathbb{E}[\varphi_z(X)e(X,Y)]\Big],$$

hence $|g_N(z) - g^*(z)| \le \|F_{\rho_N^*}|_x - F_{\rho^*}|_x\|_N + G_N(\omega)$.

Proof. Step 1: the first term is ‘⟪Φ_z, d⟫_N‘, bounded by ‘‖Φ_z‖_N ‖d‖_N ≤ ‖d‖_N‘.
Step 2: the second term is bounded by the deviation ‘G_N‘ of the residual class.
Step 3: combine through the decomposition and the tower property. □

Theorem 337. *On E'_δ, with $\Delta_N := \Delta_N(\delta/2)$ and Δ'_N as in (86),*

$$\sup_{z\in Z}\big|(R^N_{\lambda,\nu_N^*}y)(z) - (R_{\lambda,\nu^*}f)(z)\big| \le \frac{1}{\lambda}\Big[\sqrt{2\Delta_N} + \Delta'_N\Big],$$

where $R^N_{\lambda,\nu_N^*}y = -g_N/\lambda = m_{\rho_N^*}$ (ν_N^*-a.e.) and $R_{\lambda,\nu^*}f = -g^*/\lambda = m_{\rho^*}$ (ν^*-a.e.).

Proof. ‘g_N = −λ R^N y‘ and ‘g* = −λ R f‘ (‘lem:empirical-gibbs‘(a), ‘thm:m1‘(c)). □

Theorem 338. *On E'_δ,*

$$\sup_{x\in\mathcal{X}}\big|F_{\rho_N^*}(x) - F_{\rho^*}(x)\big| \le \frac{1}{\lambda}\Big[\sqrt{2\Delta_N} + \Delta'_N\Big] + \frac{2Y_{\max}}{\lambda}\sqrt{\frac{\Delta_N}{2\beta}}.$$

Proof. Step 1: ‘F_ρ_N^* = S_ν_N^*(−g_N/λ)‘ and ‘F_ρ* = S_ν*(−g*/λ)‘.

Step 2: the pointwise bounds ‘|(−g_N/λ − (−g*/λ)) φ_z| ≤ S‘ and ‘|(−g*/λ) φ_z| ≤ Ymax/λ‘.

Step 3: split ‘F_ρ_N^* − F_ρ* = ∫ (m_N − m*) φ dν_N + [∫ m* φ dν_N − ∫ m* φ dν*]‘. □

Theorem 339. *Under the hypotheses of Theorem 328, $\mathbb{P}(E'_\delta) \geq 1-\delta$: apply Theorem 328 with $\delta/2$ and Lemma 305 with $\delta/2$ to the class $\{\varphi_z e\}$, $|\varphi_z e| \leq B + Y_{\max}$, and take the intersection.*

Proof. □

Theorem 340. *On E'_δ,*

$$\|\Pi\rho_N^* - \Pi\rho^*\|_{\mathrm{TV}} \leq \frac{1}{2\lambda}\left[\sqrt{2\Delta_N} + \Delta'_N\right] + \frac{Y_{\max}}{\lambda}\sqrt{\frac{\Delta_N}{2\beta}},$$

with $\|\mu\|_{\mathrm{TV}} = \frac{1}{2}|\mu|(Z)$.

Proof. Step 1: ‘Πρ_N^* = m_N ν_N^*‘ and ‘Πρ* = m* ν*‘ with ‘m_N = −g_N/λ‘, ‘m* = −g*/λ‘.

Step 2: for every measurable ‘A‘, ‘(Πρ_N^* − Πρ*)(A) − (Πρ_N^* − Πρ*)(A^c) = ∫ σ_A (m_N − m*) dν_N^* + [∫ σ_A m* dν_N^* − ∫ σ_A m* dν*]‘.

Step 3: the two terms are bounded by ‘S = sup|m_N − m*|‘ and ‘2(Ymax/λ) ‖ν_N^* − ν*‖_TV‘. □

Theorem 341. *By Lemma 314, $\Delta_N \leq D(\delta/2)/\sqrt{N}$ and $\Delta'_N \leq D'(\delta)/\sqrt{N}$, so that on E'_δ, $\|\rho_N^* - \rho^*\|_{\mathrm{TV}} \leq \sqrt{D(\delta/2)/(2\beta\sqrt{N})}$ and $\sup_z |(R^N_{\lambda,\nu_N^*}y)(z) - (R_{\lambda,\nu^*}f)(z)| \leq \frac{1}{\lambda}[\sqrt{2D(\delta/2)/\sqrt{N}} + D'(\delta)/\sqrt{N}]$: for fixed $(\lambda,\beta,\delta,m,Y_{\max})$, (b) is $O(N^{-1/4})$ and (a), (c), (d) are $O(\beta^{-1/2}N^{-1/4})$.*

Proof. □

Theorem 342. *In the setting of Theorem 328 let $\delta \in (0,1)$, $\Delta_N := \Delta_N(\delta/2)$ and Δ'_N as in (86). With probability at least $1-\delta$ the following hold simultaneously: (a) $\|\rho_N^* - \rho^*\|_{\mathrm{TV}} \leq \sqrt{\Delta_N/(2\beta)}$ and $\|\nu_N^* - \nu^*\|_{\mathrm{TV}} \leq \sqrt{\Delta_N/(2\beta)}$; (b) $\sup_z |g_N(z) - g^*(z)| \leq \sqrt{2\Delta_N} + \Delta'_N$, i.e. $\sup_z |(R^N_{\lambda,\nu_N^*}y)(z) - (R_{\lambda,\nu^*}f)(z)| \leq \frac{1}{\lambda}[\sqrt{2\Delta_N} + \Delta'_N]$; (c) $\sup_x |F_{\rho_N^*}(x) - F_{\rho^*}(x)| \leq \frac{1}{\lambda}[\sqrt{2\Delta_N} + \Delta'_N] + \frac{2Y_{\max}}{\lambda}\sqrt{\Delta_N/(2\beta)}$; (d) $\|\Pi\rho_N^* - \Pi\rho^*\|_{\mathrm{TV}} \leq \frac{1}{2\lambda}[\sqrt{2\Delta_N} + \Delta'_N] + \frac{Y_{\max}}{\lambda}\sqrt{\Delta_N/(2\beta)}$.*

Proof. □

5.5 Improved rate via linearization

Theorem 343. *For $m_i \in L^1(\nu_i)$, $m_1\nu_1 - m_2\nu_2 = \psi\,(\nu_1+\nu_2)$ with ψ the mixed density.*

Proof. □

Theorem 344. *For $gm_i \in L^1(\nu_i)$, $\int m_1 g\,\mathrm{d}\nu_1 - \int m_2 g\,\mathrm{d}\nu_2 = \int \psi g\,\mathrm{d}(\nu_1+\nu_2)$.*

Proof. □

Theorem 345. *For bounded measurable m_1, m_2, $|m_1\nu_1 - m_2\nu_2|(\alpha) = \int |\psi|\,\mathrm{d}(\nu_1+\nu_2)$.*

Proof. □

Theorem 346. *For bounded measurable m_1, m_2 with $|m_2| \leq M$, $\int |\psi|\,\mathrm{d}(\nu_1+\nu_2) \leq \int |m_1 - m_2|\,\mathrm{d}\nu_1 + 2M\|\nu_1 - \nu_2\|_{\mathrm{TV}}$, from $|w_1m_1 - w_2m_2| \leq w_1|m_1 - m_2| + |w_1 - w_2||m_2|$.*

Proof. Step 1: the pointwise bound '$|w_1\, m_1 - w_2\, m_2| \le w_1\, |m_1 - m_2| + M\, |w_1 - w_2|$'.
Step 2: integrate, using '$\int w_1\, g\, d\mu = \int g\, d\nu_1$' and 'lem:rn-deriv-sub-tv'. ☐

Theorem 347. *If* $\|f\|_{L^2(P_X)} \le Y_{\max}$ *then* $\sup_z |m^*(z)| \le M_* = Y_{\max}/\lambda$.

Proof. ☐

Theorem 348. $F_{\rho^*}(x) = \int_Z m^*(z)\varphi_z(x)\, \nu^*(\,\mathrm{d}z)$ *for every* x.

Proof. ☐

Theorem 349. *If* $|Y| \le Y_{\max}$ *a.s. and* $\|f\| \le Y_{\max}$*, then* $|e| \le E_* = M_* + Y_{\max}$ *a.s.*

Proof. ☐

Theorem 350. *If* $|y_i| \le Y_{\max}$ *for all* i *then* $\sup_z |m_N(z)| \le Y_{\max}/\lambda$.

Proof. ☐

Theorem 351. $F_{\rho^*_N}(x) = \int_Z m_N(z)\varphi_z(x)\, \nu^*_N(\,\mathrm{d}z)$ *for every* x.

Proof. ☐

Theorem 352. $\Pi\rho^*_N = m_N\, \nu^*_N$.

Proof. ☐

Theorem 353. $h(x) = \int_Z \varphi_z(x)\, \Pi\varsigma(\,\mathrm{d}z) = \int_Z \psi(z)\varphi_z(x)\, \mu(\,\mathrm{d}z)$ *for every* x*, where* $\Pi\varsigma = m_N\nu^*_N - m^*\nu^* = \psi\mu$.

Proof. ☐

Theorem 354. $D_\varsigma = |\Pi\varsigma|(Z) = \int_Z |\psi|\, \mathrm{d}\mu$.

Proof. ☐

Theorem 355. *Let* $\Psi = \{\Psi_i : i \in I\}$ *be a class of measurable functions on* $\mathcal{X} \times \mathbb{R}$ *bounded by one,* κ *a finite measure on* I*,* $c \in L^1(\kappa)$ *and* $e \in L^1(P)$*. Then, with* $H(p) := \int_I c(i)\Psi_i(p)\, \kappa(\,\mathrm{d}i)$*,* $(P_N - P)[He] = \int_I c(i)\, (P_N - P)(\Psi_i e)\, \kappa(\,\mathrm{d}i)$ *(Fubini, since* $\iint |c(i)\Psi_i(p)e(p)|\, \kappa(\,\mathrm{d}i)P(\,\mathrm{d}p) \le \|c\|_{L^1(\kappa)}\|e\|_{L^1(P)} < \infty$*).*

Proof. Step 1: the empirical mean, a finite sum.
Step 2: the population mean, by Fubini. ☐

Theorem 356. *Under the hypotheses of Lemma 355,* $|(P_N - P)[He]| \le \|c\|_{L^1(\kappa)} \sup_{i\in I} |(P_N - P)(\Psi_i e)|$.

Proof. ☐

Theorem 357. *Under the assumptions of Theorem 328, for every realization of the sample, deterministically,*

$$\tfrac{1}{2}\|h\|_N^2 + \tfrac{1}{2}\|h\|^2_{L^2(P_X)} + \beta\,\mathcal{K}_N = -(P_N - P)[\ell_{F_{\rho^*_N}} - \ell_{F_{\rho^*}}] = -(P_N - P)[he] - \tfrac{1}{2}(P_N - P)[h^2].$$

(Add the two quadratic expansions of Lemma **??***; the entropy terms* $\beta\,\mathrm{KL}(\cdot\|\mu_U)$ *cancel, and* $\ell_{F_{\rho^*_N}} - \ell_{F_{\rho^*}} = he + \frac{1}{2}h^2$*.)*

Proof. Step 1: ‘‖F_ρ_N^* − F_ρ*‖²_L²(P_X) = ∫ h² dP_X‘.
Step 2: the risks are the means of the square loss.
Step 3: add the two expansions; the entropy terms cancel. □

Theorem 358. $\frac{1}{2}\|h\|_N^2 + \frac{1}{2}\|h\|^2 + \beta\mathcal{K}_N = -(P_N - P)[he] - \frac{1}{2}(P_N - P)[h^2]$, *since* $\ell_{F_{\rho_N^*}} - \ell_{F_{\rho^*}} = \frac{1}{2}(h+e)^2 - \frac{1}{2}e^2 = he + \frac{1}{2}h^2$.

Proof. □

Theorem 359. *For each direction separately,* $\frac{1}{2}\|h\|_N^2 + \beta\,\mathrm{KL}(\rho^*\|\rho_N^*) \le -(P_N - P)[he + \frac{1}{2}h^2]$ *and* $\frac{1}{2}\|h\|^2 + \beta\,\mathrm{KL}(\rho_N^*\|\rho^*) \le -(P_N - P)[he + \frac{1}{2}h^2]$ *(drop one of the two nonnegative expansions).*

Proof. □

Theorem 360. $(P_N - P)[he] = \int_Z [(P_N - P)(\varphi_z e)]\,\Pi\varsigma(\,\mathrm{d}z)$ *and* $|(P_N - P)[he]| \le D_\varsigma G_N$.

Proof. □

Theorem 361. $(P_N - P)[h^2] = \iint_{Z\times Z}[(P_N - P)(\varphi_z\varphi_{z'})]\,\Pi\varsigma(\,\mathrm{d}z)\Pi\varsigma(\,\mathrm{d}z')$ *and* $|(P_N - P)[h^2]| \le D_\varsigma^2 H_N$.

Proof. Step 1: ‘h(x)² = ∬ ψ(z)ψ(z’) φ_z(x)φ_z’(x) μ(dz)μ(dz’)‘.
Step 2: ‘∫ |ψ(z)ψ(z’)| μ(dz)μ(dz’) = D_ς²‘. □

Theorem 362. *Deterministically,* $\frac{1}{2}\|h\|_N^2 + \frac{1}{2}\|h\|^2 + \beta\mathcal{K}_N \le D_\varsigma G_N + \frac{1}{2}D_\varsigma^2 H_N$.

Proof. □

Theorem 363. $g^*(z) = \langle\varphi_z, F_{\rho^*} - f\rangle_{L^2(P_X)} = \mathbb{E}[\varphi_z(X)(F_{\rho^*}(X) - Y)] = P(\varphi_z e)$, *since* $\mathbb{E}[(Y - f(X))\varphi_z(X)] = 0$.

Proof. Step 1: replace the ‘L²‘ residual by ‘F_ρ* − f‘.
Step 2: pull back to ‘Q‘.
Step 3: the tower property ‘∫ (y − f(x)) φ_z(x) dQ = 0‘. □

Theorem 364. *For every* z, $\lambda(m_N - m^*)(z) = -P_N(\varphi_z h) - (P_N - P)(\varphi_z e)$.

Proof. □

Theorem 365. $\sup_z |m_N(z) - m^*(z)| \le \frac{1}{\lambda}[\|h\|_N + G_N]$, *because* $|P_N(\varphi_z h)| \le \|\Phi_z\|_N\|h\|_N \le \|h\|_N$.

Proof. □

Theorem 366. *Under Theorem* **??**(c) *and Lemma* **??**(a), *if* $\sup_z |m^*(z)| \le M_1$ *then, deterministically,* $D_\varsigma \le \frac{1}{\lambda}[\|h\|_N + G_N] + M_1\sqrt{\mathcal{K}_N}$ *(Lemma 367 with* M_* *replaced by* M_1*; the same proof).*

Proof. Step 1: ‘D_ς = ∫ |ψ| dμ ≤ ∫ |m_N − m*| dν_N^* + M_* · 2‖ν_N^* − ν*‖_TV‘.
Step 2: ‘∫ |m_N − m*| dν_N^* ≤ sup |m_N − m*| ≤ C‘.
Step 3: ‘2‖ν_N^* − ν*‖_TV ≤ 2‖ρ_N^* − ρ*‖_TV ≤ √𝒦_N‘. □

Theorem 367. *Under Theorem* **??**(c), (d) *and Lemma* **??**(a), *deterministically,*

$$D_\varsigma = |\Pi\rho_N^* - \Pi\rho^*|(Z) \le \sup_z |m_N - m^*| + M_*\,|\nu_N^* - \nu^*|(Z) \le \frac{1}{\lambda}[\|h\|_N + G_N] + M_*\sqrt{\mathcal{K}_N}.$$

($\Pi\rho_N^* - \Pi\rho^* = (m_N - m^*)\nu_N^* + m^*(\nu_N^* - \nu^*)$*, subadditivity of the total variation,* $|\nu_N^* - \nu^*|(Z) = 2\|\nu_N^* - \nu^*\|_{\mathrm{TV}} \le 2\|\rho_N^* - \rho^*\|_{\mathrm{TV}} \le \sqrt{\mathcal{K}_N}$ *by Lemma 852.)*

Proof. □

Theorem 368. *Let $u, u_P, v, D, G, H \geq 0$ and $c \geq 1$ with $c^2H \leq \frac{1}{8}$, $\frac{1}{2}u^2 + \frac{1}{2}u_P^2 + v^2 \leq DG + \frac{1}{2}D^2H$ and $D \leq c(u+v+G)$. Then $u+v \leq 9cG$, $u_P \leq 6cG$ and $D \leq 10c^2G$. (With $s = u+v$: $\frac{1}{4}s^2 \leq \frac{1}{2}u^2+v^2$, $D^2H \leq \frac{1}{8}(s+G)^2$, hence $s^2 \leq 8cGs + 9cG^2$ and $s \leq 9cG$.)*

Proof. Step 1: 'DG ≤ c(s + G)G' and 'D²H ≤ (1/8)(s + G)²'.
Step 2: the quadratic inequality 's² ≤ 8cGs + 9cG²'.
Step 3: solve the quadratic inequality.
'u_P² ≤ 2DG + D²H ≤ 20c²G² + (1/8)(10cG)² ≤ 36c²G²'. □

Theorem 369. *Assume the setting of Theorem 370 and $\sup_z |m^*(z)| \leq M_1$; put $c_1(M_1) := \max\{1, 1/\lambda, M_1/\sqrt{\beta}\}$. For any upper bounds $\eta' \geq G_N$, $\eta'' \geq H_N$ with $c_1(M_1)^2\eta'' \leq \frac{1}{8}$: $\|h\|_N + \sqrt{\beta\mathcal{K}_N} \leq 9c_1(M_1)\eta'$, $\|h\| \leq 6c_1(M_1)\eta'$, $D_\varsigma \leq 10c_1(M_1)^2\eta'$, $\mathcal{K}_N \leq \frac{81c_1(M_1)^2}{\beta}\eta'^2$ and $\|\rho_N^* - \rho^*\|_{\mathrm{TV}} \leq \frac{9c_1(M_1)}{2\sqrt{\beta}}\eta'$ (the proof of Theorem 370 with Lemma 366).*

Proof. Step 1: the self-bounding inequality with 'u_P² = ‖h‖²' and 'v² = β𝒦_N'.
Step 2: 'lem:coefficient-tv' in the form 'D ≤ c(u + v + G)'.
Step 3: the arithmetic closure.
(b): '𝒦_N = v²/β ≤ (9 c η')²/β'.
(b): '‖ρ_N^* − ρ*‖_TV ≤ ½ √𝒦_N = v/(2 √β)'. □

Theorem 370. *Assume the setting of Theorem 328. Put $c_1 := \max\{1, 1/\lambda, M_*/\sqrt{\beta}\}$ and $N_B(\delta) := \lceil 64c_1^4D''(\delta)^2\rceil$ ($\Leftrightarrow c_1^2\eta_N'' \leq \frac{1}{8}$). Let $\delta \in (0,1)$ and $N \geq N_B(\delta)$. On the event Ω_δ the following hold simultaneously.*

(a) $\|h\|_N + \sqrt{\beta\mathcal{K}_N} \leq 9c_1\eta_N'$, $\|h\|_{L^2(P_X)} \leq 6c_1\eta_N'$, $D_\varsigma \leq 10c_1^2\eta_N'$.

(b) $\mathcal{K}_N \leq \frac{81c_1^2}{\beta}\eta_N'^2$ *and* $\|\rho_N^* - \rho^*\|_{\mathrm{TV}} \leq \frac{9c_1}{2\sqrt{\beta}}\eta_N'$.

(In Lean the statement is deterministic: for any upper bounds $\eta' \geq G_N$, $\eta'' \geq H_N$ with $c_1^2\eta'' \leq \frac{1}{8}$.)

Proof. □

Theorem 371. *Assume the setting of Theorem 369. For any upper bounds $\eta' \geq G_N$, $\eta'' \geq H_N$ with $c_1(M_1)^2\eta'' \leq \frac{1}{8}$: $\|\rho_N^*-\rho^*\|_{\mathrm{TV}}, \|\nu_N^*-\nu^*\|_{\mathrm{TV}} \leq \frac{9c_1(M_1)}{2\sqrt{\beta}}\eta'$, $\sup_z |m_N - m^*| \leq \frac{1}{\lambda}[\|h\|_N + \eta'] \leq \frac{10c_1(M_1)}{\lambda}\eta'$, $\sup_x |h| \leq D_\varsigma \leq 10c_1(M_1)^2\eta'$ and $\|\Pi\rho_N^* - \Pi\rho^*\|_{\mathrm{TV}} = \frac{1}{2}D_\varsigma \leq 5c_1(M_1)^2\eta'$.*

Proof. (c): '|h(x)| = |∫ ψ φ_z(x) dμ| ≤ ∫ |ψ| dμ = D_ς'. □

Theorem 372. *Assume the setting of Theorem 370, $\delta \in (0,1)$ and $N \geq N_B(\delta)$. On the event Ω_δ the following hold simultaneously.*

(a) $\|\rho_N^* - \rho^*\|_{\mathrm{TV}} \leq \frac{9c_1}{2\sqrt{\beta}}\eta_N'$ *and* $\|\nu_N^* - \nu^*\|_{\mathrm{TV}} \leq \frac{9c_1}{2\sqrt{\beta}}\eta_N'$.

(b) $\sup_z |(R^N_{\lambda,\nu_N^*}y)(z) - (R_{\lambda,\nu^*}f)(z)| = \sup_z |m_N(z) - m^*(z)| \leq \frac{1}{\lambda}[\|h\|_N + \eta_N'] \leq \frac{10c_1}{\lambda}\eta_N'$.

(c) $\sup_x |F_{\rho_N^*}(x) - F_{\rho^*}(x)| \leq D_\varsigma \leq 10c_1^2\eta_N'$, *since* $|h(x)| = |\int \varphi_z(x)\,\mathrm{d}\Pi\varsigma| \leq D_\varsigma$.

(d) $\|\Pi\rho_N^* - \Pi\rho^*\|_{\mathrm{TV}} = \frac{1}{2}D_\varsigma \leq 5c_1^2\eta_N'$.

Proof. □

Theorem 373. *Fix $\delta \in (0,1)$ and apply Lemma 305 with $\delta_1 = \delta/2$ and Lemma 303 with $\delta_2 = \delta/2$: with $\eta'_N = 2E_*\mathfrak{R}_N(\Phi) + E_*\sqrt{2\log(4/\delta)/N}$ and $\eta''_N = 2\mathfrak{R}_N(\Psi) + \sqrt{2\log(4/\delta)/N}$, $\mathbb{P}(\Omega_\delta) = \mathbb{P}(\{G_N \le \eta'_N\} \cap \{H_N \le \eta''_N\}) \ge 1 - \delta$.*

Proof. Step 1: ‘lem:GN‘ at level ‘δ/2‘ for ‘e = F_ρ* − y‘.
Step 2: ‘lem:HN‘ at level ‘δ/2‘.
Step 3: ‘A‘ is null-measurable (a.e. equal to the measurable event of the clipped residual) and ‘B‘ is measurable.
Step 4: the union bound. □

Theorem 374. *If $|e| \le E$ Q-a.s. for some $E > 0$, then with $\eta'_N = 2E\mathfrak{R}_N(\Phi) + E\sqrt{2\log(4/\delta)/N}$ and $\eta''_N = 2\mathfrak{R}_N(\Psi) + \sqrt{2\log(4/\delta)/N}$, $\mathbb{P}(\{G_N \le \eta'_N\} \cap \{H_N \le \eta''_N\}) \ge 1 - \delta$ (Lemmas 305 and 303 at level $\delta/2$ each, and a union bound).*

Proof. Step 1: ‘lem:GN‘ at level ‘δ/2‘ for ‘e = F_ρ* − y‘.
Step 2: ‘lem:HN‘ at level ‘δ/2‘.
Step 3: ‘A‘ is null-measurable (a.e. equal to the measurable event of the clipped residual) and ‘B‘ is measurable.
Step 4: the union bound. □

Theorem 375. *If $|Y| \le Y_{\max}$ a.s. and $\sup_z |m^*(z)| \le M_1$, then $|e| \le M_1 + Y_{\max}$ a.s. ($|F_{\rho^*}| = |S_{\nu^*} m^*| \le M_1$).*

Proof. □

Theorem 376. *If $\sup_z |m^*(z)| \le M_1$, then with $E_* := M_1 + Y_{\max}$ in η'_N, $\mathbb{P}(\Omega_\delta) \ge 1 - \delta$ (Lemma 374 with Lemma 375).*

Proof. □

Theorem 377. *If $\mathfrak{R}_N(\Phi) \le C_{\mathrm{P}}\sqrt{(m+1)/N}$ then $\eta'_N \le E_* D'(\delta)/\sqrt{N}$, $D'(\delta) = 2C_{\mathrm{P}}\sqrt{m+1} + \sqrt{2\log(4/\delta)}$.*

Proof. □

Theorem 378. *If $\mathfrak{R}_N(\Psi) \le C'_{\mathrm{P}}\sqrt{(m+1)/N}$ then $\eta''_N \le D''(\delta)/\sqrt{N}$, $D''(\delta) = 2C'_{\mathrm{P}}\sqrt{m+1} + \sqrt{2\log(4/\delta)}$.*

Proof. □

Theorem 379. *For $c \ge 0$ and $D'' \ge 0$, $N \ge 64c^4 D''^2$ implies $c^2 D''/\sqrt{N} \le \frac{1}{8}$; this is the condition $N \ge N_B(\delta) = \lceil 64c_1^4 D''(\delta)^2 \rceil$ of Theorem 370.*

Proof. □

Theorem 380. *Assume the setting of Theorem 369, the complexity bounds $\mathfrak{R}_N(\Phi) \le C_{\mathrm{P}}\sqrt{(m+1)/N}$, $\mathfrak{R}_N(\Psi) \le C'_{\mathrm{P}}\sqrt{(m+1)/N}$ and $N \ge 64c_1(M_1)^4 D''(\delta)^2$. On the event Ω_δ (with $E := M_1 + Y_{\max}$ in η'_N), with $D'(\delta) = 2C_{\mathrm{P}}\sqrt{m+1} + \sqrt{2\log(4/\delta)}$: $\mathcal{K}_N \le \frac{121c_1(M_1)^2E^2D'(\delta)^2}{\beta N}$, $\|h\|_N, \|h\| \le \frac{11c_1(M_1)ED'(\delta)}{\sqrt{N}}$, $\|\rho^*_N - \rho^*\|_{\mathrm{TV}} \le \frac{11c_1(M_1)ED'(\delta)}{2\sqrt{\beta N}}$, $\sup_z |m_N - m^*| \le \frac{12c_1(M_1)ED'(\delta)}{\lambda\sqrt{N}}$ and $\sup_x |F_{\rho^*_N} - F_{\rho^*}| \le \frac{12c_1(M_1)^2ED'(\delta)}{\sqrt{N}}$.*

Proof. Step 1: the threshold ‘N ≥ N_B(δ)‘ gives ‘c_1² η”_N ≤ 1/8‘.
Step 2: ‘η’_N ≤ E_* D’(δ)/√N‘.
Step 3: the constants ‘9, 6, 10, 81‘ of ‘thm:m3-double-prime‘ are rounded to ‘11, 11, 12, 121‘. □

Theorem 381. *On Ω_δ, for $N \geq N_B(\delta)$ and under Lemmas 314 and 315: with $D'(\delta) = 2C_{\mathrm{P}}\sqrt{m+1} + \sqrt{2\log(4/\delta)}$, $\mathcal{K}_N \leq \frac{121c_1^2E_*^2D'(\delta)^2}{\beta N} = O(1/N)$, $\|h\|_N, \|h\| \leq \frac{11c_1E_*D'(\delta)}{\sqrt{N}}$, $\|\rho_N^* - \rho^*\|_{\mathrm{TV}} \leq \frac{11c_1E_*D'(\delta)}{2\sqrt{\beta N}}$, $\sup_z |m_N - m^*| \leq \frac{12c_1E_*D'(\delta)}{\lambda\sqrt{N}}$ and $\sup_x |F_{\rho_N^*} - F_{\rho^*}| \leq \frac{12c_1^2E_*D'(\delta)}{\sqrt{N}}$, all of order $N^{-1/2}$ with no logarithm.*

Proof. □

Chapter 6

Regularization limits and the geometry of the limit

Section 6 of the paper: representability, the canonical (minimum-norm) ridgelet transform, the energy comparison with a competitor built on the reference measure (`thm:m4-energy`), the convergence of the conditional mean amplitude to the canonical ridgelet transform as $\lambda \to 0$ and $\kappa = \lambda/\beta \to 0$, the order of the limits, and the reduction of the zero-temperature problem to total-variation regularization.

6.1 Reference measure and representability

Theorem 382. *Assume (R).* $u_0^\dagger := P_{(\ker S_{\nu_0})^\perp} u_0$ *is a representer and does not depend on the choice of the representer* u_0*: for every representer* u_0 *of* f*,* $S_{\nu_0}^\dagger f = P_{(\ker S_{\nu_0})^\perp} u_0$ *and* $S_{\nu_0} u_0^\dagger = f$*. (Two representers differ by an element of* $\ker S_{\nu_0}$*, so their projections onto* $(\ker S_{\nu_0})^\perp$ *coincide; and* $u_0 - u_0^\dagger \in \ker S_{\nu_0}$*.)*

Proof. The chosen representer and 'u$_0$' have the same synthesis 'f'. ☐

Theorem 383. *Assume (R). For every representer* u *of* f*,* $\|u\|_{L^2(\nu_0)}^2 = \|u_0^\dagger\|_{L^2(\nu_0)}^2 + \|u - u_0^\dagger\|_{L^2(\nu_0)}^2$*. In particular* $\|u_0^\dagger\| \leq \|u\|$*, with equality if and only if* $u = u_0^\dagger$*. (*$u - u_0^\dagger \in \ker S_{\nu_0} \perp u_0^\dagger$*, and Pythagoras.)*

Proof. Pythagoras for the orthogonal projections onto 'K = (ker S)$^\perp$' and 'K$^\perp$', where the projection onto 'K$^\perp$' is 'u − P_K u'. ☐

Theorem 384. *Assume (R).* $u_0^\dagger$ *is the unique representer in* $(\ker S_{\nu_0})^\perp = \overline{\operatorname{ran} S_{\nu_0}^*}$*, and if* $h \in L^2(\nu_0)$ *satisfies* $S_{\nu_0} h = f$ *and* $\|h\|_{L^2(\nu_0)} \leq \|u_0^\dagger\|_{L^2(\nu_0)}$ *then* $h = u_0^\dagger$*. (A representer* $h \in (\ker S_{\nu_0})^\perp$ *satisfies* $h - u_0^\dagger \in \ker S_{\nu_0} \cap (\ker S_{\nu_0})^\perp = \{0\}$*; for the last claim, Pythagoras with* $\|h\|^2 \leq \|u_0^\dagger\|^2$ *forces* $\|h - u_0^\dagger\| = 0$*.)*

Proof. ☐

6.2 Energy comparison

Theorem 385. *For* $\nu \in \mathcal{P}(Z)$ *and* $u \in L^2(\nu)$ *measurable,* $\int a^2 \, d\rho_{\nu,u} = \|u\|_{L^2(\nu)}^2 + \sigma^2$ *(*$\mathbb{E}X^2 = \mu^2 + \sigma^2$ *for* $X \sim \mathcal{N}(\mu, \sigma^2)$*, integrated against* ν*).*

Proof. □

Theorem 386. *For $\nu \in \mathcal{P}(Z)$ and $u \in L^2(\nu)$ measurable, $\int |a| \,\mathrm{d}\rho_{\nu,u} \le \|u\|_{L^1(\nu)} + \sigma\sqrt{2/\pi} < \infty$ ($\mathbb{E}|X| \le |\mu| + \sigma\sqrt{2/\pi}$ for $X \sim \mathcal{N}(\mu, \sigma^2)$, and $u \in L^2(\nu) \subset L^1(\nu)$).*

Proof. Step 1: the inner Gaussian moment bound '∫ |a| 𝒩(u(z), σ²)(da) ≤ |u(z)| + σ √(2/π)'.

Step 2: integrate against 'ν'. □

Theorem 387. *For $\nu \in \mathcal{P}(Z)$ and $u \in L^2(\nu)$ measurable, $F_{\rho_{\nu,u}}(x) = \int_Z [\int_{\mathbb{R}} a\, \mathcal{N}(u(z), \sigma^2)(\mathrm{d}a)] \varphi_z(x)\, \nu(\mathrm{d}z) = (S_\nu u)(x)$ for every x (Fubini, since $\int |a| \,\mathrm{d}\rho_{\nu,u} < \infty$).*

Proof. □

Theorem 388. *For $\nu \in \mathcal{P}(Z)$ and $u \in L^2(\nu)$ (measurable), $\int |a| \,\mathrm{d}\rho_{\nu,u} \le \|u\|_{L^1(\nu)} + \sigma\sqrt{2/\pi} < \infty$, $\int a^2 \,\mathrm{d}\rho_{\nu,u} = \|u\|^2_{L^2(\nu)} + \sigma^2$, $F_{\rho_{\nu,u}} = S_\nu u$ and $\Pi\rho_{\nu,u} = u\,\nu$. (For $X \sim \mathcal{N}(\mu, \sigma^2)$, $\mathbb{E}|X| \le |\mu| + \sigma\sqrt{2/\pi}$ and $\mathbb{E}X^2 = \mu^2 + \sigma^2$; apply this with $\mu = u(z)$ and integrate against ν; Fubini gives $F_{\rho_{\nu,u}}(x) = \int_Z [\int_{\mathbb{R}} a\, \mathcal{N}(u(z), \sigma^2)(\mathrm{d}a)] \varphi_z(x) \nu(\mathrm{d}z) = (S_\nu u)(x)$ and likewise $\Pi\rho_{\nu,u} = u\nu$.)*

Proof. Step 1: transport the set integral through the disintegration.
Step 2: the inner integral is the Gaussian mean 'u(z)'. □

Theorem 389. *If $\nu \ll \nu'$ and u, u' are measurable then $\rho_{\nu,u} \ll \rho_{\nu',u'}$ with $\frac{\mathrm{d}\rho_{\nu,u}}{\mathrm{d}\rho_{\nu',u'}}(a, z) = \frac{\mathrm{d}\nu}{\mathrm{d}\nu'}(z) \exp\Big(\frac{-(a-u(z))^2 + (a-u'(z))^2}{2\sigma^2}\Big)$ $\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}(z) \exp\Big(\kappa(a(u-u')(z) - \frac{1}{2}(u^2 - u'^2)(z))\Big)$.*

Proof. The exponent in the Gaussian shift identity, with 'σ² = β/λ'.
For fixed 'z', the Gaussian shift identity 'E(·, z) 𝒩(u'(z), σ²) = 𝒩(u(z), σ²)'. □

Theorem 390. *If $\nu, \nu' \in \mathcal{P}(Z)$, $u \in L^2(\nu)$, u' measurable, $\nu \ll \nu'$, $\mathrm{KL}(\nu\|\nu') < \infty$ and $u' \in L^2(\nu)$, then $\mathrm{KL}(\rho_{\nu,u}\|\rho_{\nu',u'}) < \infty$ and*

$$\mathrm{KL}(\rho_{\nu,u}\|\rho_{\nu',u'}) = \mathrm{KL}(\nu\|\nu') + \frac{\kappa}{2}\|u - u'\|^2_{L^2(\nu)}.$$

(By Lemma 389, $\log \frac{\mathrm{d}\rho_{\nu,u}}{\mathrm{d}\rho_{\nu',u'}} = \log \frac{\mathrm{d}\nu}{\mathrm{d}\nu'}(z) + \kappa(a(u-u') - \frac{1}{2}(u^2 - u'^2))$, which is $\rho_{\nu,u}$-integrable since $|a(u-u')| \le \frac{1}{2}(a^2 + (u-u')^2)$ with $\int a^2 \,\mathrm{d}\rho_{\nu,u} < \infty$; and $\int a\, \mathcal{N}(u, \sigma^2)(\mathrm{d}a) = u$ gives $\int \kappa(u(u - u') - \frac{1}{2}(u^2 - u'^2))\,\mathrm{d}\nu = \frac{\kappa}{2}\|u - u'\|^2_{L^2(\nu)}$.)

Proof. Step 1: the log-likelihood ratio, 'ρ'-a.e. (where 'dν/dν'' ∈ (0, ∞)').
Step 2: integrability of the log-likelihood ratio.
Step 3: integrate, using Fubini through the disintegration for the term 'a (u − u')'. □

Theorem 391. *If $\nu \in \mathcal{P}(Z)$, $u \in L^2(\nu)$ measurable, $\nu \ll \nu_0$ and $\mathrm{KL}(\nu\|\nu_0) < \infty$, then $\rho_{\nu,u} \in \mathcal{D}$ and*

$$\mathrm{KL}(\rho_{\nu,u}\|\mu_U) = \mathrm{KL}(\nu\|\nu_0) + \frac{\lambda}{2\beta}\|u\|^2_{L^2(\nu)} = \mathrm{KL}(\nu\|\nu_0) + \frac{\kappa}{2}\|u\|^2_{L^2(\nu)}.$$

(Lemma 390 with $\nu' = \nu_0$ and $u' = 0$, since $\mu_U = \rho_{\nu_0,0}$.)

Proof. □

Theorem 392. *Let $\nu_0, \nu \in \mathcal{P}(Z)$ with $\nu = w\,\nu_0$ for a measurable w with $e^{-c} \le w \le e^c$. Then $\mathrm{KL}(\nu\|\nu_0) < \infty$ and $\mathrm{KL}(\nu_0\|\nu) < \infty$, since $\log \frac{\mathrm{d}\nu}{\mathrm{d}\nu_0} = \log w$ and $\log \frac{\mathrm{d}\nu_0}{\mathrm{d}\nu} = -\log w$ are bounded by $|c|$.*

Proof. 'log w' is bounded by '|c|', hence integrable for every probability measure.
'llr ν ν_0 = log w' and 'llr ν_0 ν = −log w'. ☐

Theorem 393. *Let $u_0 \in L^2(\nu_0)$ be a representer of f and $\tilde\rho := \tilde\rho_{u_0} = \rho_{\nu_0,u_0}$. Then $\tilde\rho \in \mathcal{D}$, $F_{\tilde\rho} = S_{\nu_0}u_0 = f$, $L(\tilde\rho) = 0$, $\mathrm{KL}(\tilde\rho\|\mu_U) = \frac{\lambda}{2\beta}\|u_0\|^2_{L^2(\nu_0)}$ and $\mathcal{F}(\tilde\rho) = \frac{\lambda}{2}\|u_0\|^2_{L^2(\nu_0)}$ (Lemma 391 with $\nu = \nu_0$).*

Proof. Step 1: 'KL(ρ̃‖μ_U) = (κ/2)‖u$_0$‖²'.
Step 2: 'F_ρ̃ = S_ν0 u$_0$ = f', so 'L(ρ̃) = 0'. ☐

Theorem 394. *$\rho^* = \rho_{\nu^*,m^*}$ with $m^* = -s^*/\lambda$ and variance β/λ (Theorem 225).*

Proof. ☐

Theorem 395. *$\mathrm{KL}(\nu^*\|\nu_0) < \infty$ and $\mathrm{KL}(\nu_0\|\nu^*) < \infty$: by (**??**), $\log\frac{\mathrm{d}\nu^*}{\mathrm{d}\nu_0}$ is bounded.*

Proof. ☐

Theorem 396. Chain rule.

$$\mathrm{KL}(\rho^*\|\mu_U) = \mathrm{KL}(\nu^*\|\nu_0) + \frac{\lambda}{2\beta}\|m^*\|^2_{L^2(\nu^*)},$$

and consequently, under (R),

$$\|m^*\|^2_{L^2(\nu^*)} + \frac{2\beta}{\lambda}\,\mathrm{KL}(\nu^*\|\nu_0) + \frac{2}{\lambda}L(\rho^*) \le \|u_0^\dagger\|^2_{L^2(\nu_0)}.$$

(By Theorem 225, $\rho^* = \rho_{\nu^*,m^*}$; $\nu^* \ll \nu_0$ with $\mathrm{KL}(\nu^*\|\nu_0) < \infty$ and $m^* \in L^2(\nu^*)$, so Lemma 391 applies; then substitute into Theorem 397.)

Proof. ☐

Theorem 397. Energy comparison. *Assume (A1), (A3), (A4), (A5), $f \in L^2(P_X)$ and (R). Let u_0 be any representer of f and $u_0^\dagger$ the minimum-norm representer. Then*

$$L(\rho^*) + \beta\ \mathrm{KL}(\rho^*\|\mu_U) \le \frac{\lambda}{2}\|u_0\|^2_{L^2(\nu_0)}, \qquad \textit{hence} \qquad L(\rho^*) + \beta\ \mathrm{KL}(\rho^*\|\mu_U) \le \frac{\lambda}{2}\|u_0^\dagger\|^2_{L^2(\nu_0)}.$$

(Minimality of ρ^* against the competitor $\tilde\rho_{u_0}$, Lemma 393.)

Proof. ☐

Theorem 398. *Under the hypotheses of Theorem 397,*

$$\|r^*\|^2_{L^2(P_X)} = 2L(\rho^*) \le \lambda\|u_0^\dagger\|^2, \qquad \mathrm{KL}(\rho^*\|\mu_U) \le \frac{\kappa}{2}\|u_0^\dagger\|^2, \qquad \mathrm{KL}(\nu^*\|\nu_0) \le \frac{\kappa}{2}\|u_0^\dagger\|^2.$$

(The first two from $L(\rho^*) = \frac{1}{2}\|r^*\|^2$, $\mathrm{KL} \ge 0$ and $L \ge 0$; the third from the chain rule of Theorem 396 and $\|m^*\|^2 \ge 0$.)

Proof. ☐

Theorem 399. Pythagorean identity. *For every representer* u_0*,*

$$\|u_0\|^2_{L^2(\nu_0)} = \frac{4}{\lambda} L(\rho^*) + \|m^*\|^2_{L^2(\nu^*)} + \|u_0 - m^*\|^2_{L^2(\nu_0)} + \frac{2}{\kappa}\Big[\mathrm{KL}(\nu^*\|\nu_0) + \mathrm{KL}(\nu_0\|\nu^*)\Big].$$

(Apply the strong-convexity identity of Lemma 212 to the competitor $\tilde\rho = \tilde\rho_{u_0}$*:* $\mathcal{F}(\tilde\rho) - \mathcal{F}(\rho^*) = L(\rho^*) + \beta\,\mathrm{KL}(\tilde\rho\|\rho^*)$*, so* $\frac{\lambda}{2}\|u_0\|^2 = 2L(\rho^*) + \beta\,\mathrm{KL}(\rho^*\|\mu_U) + \beta\,\mathrm{KL}(\tilde\rho\|\rho^*)$*; then Theorem 396 and Lemma 390 with* $\nu = \nu_0$*,* $u = u_0$*,* $\nu' = \nu^*$*,* $u' = m^*$ *give* $\mathrm{KL}(\tilde\rho\|\rho^*) = \mathrm{KL}(\nu_0\|\nu^*) + \frac{\kappa}{2}\|u_0 - m^*\|^2_{L^2(\nu_0)}$*; multiply by* $2/\lambda$*.)*

Proof. Step 1: the strong-convexity identity at the competitor.
Step 2: ‘KL(ρ*‖μ_U)‘ by the chain rule and ‘KL(ρ̃‖ρ*)‘ by the pair formula.
Step 3: rearrange and multiply by ‘2/λ‘. □

Theorem 400. *Since every term on the right-hand side of the Pythagorean identity is nonnegative, with* $u_0 = u_0^\dagger$*,*

$$L(\rho^*) \le \frac{\lambda}{4}\|u_0^\dagger\|^2, \qquad \|u_0^\dagger - m^*\|_{L^2(\nu_0)} \le \|u_0^\dagger\|, \qquad \mathrm{KL}(\nu^*\|\nu_0) + \mathrm{KL}(\nu_0\|\nu^*) \le \frac{\kappa}{2}\|u_0^\dagger\|^2.$$

Proof. □

Theorem 401. Normalizing constant. *With* Z_**,* Z_0 *as in Theorem 226 and* $\tilde Z := Z_*/Z_0 = \int e^{\kappa m^{*2}/2}\,\mathrm{d}\nu_0$*,*

$$\frac{\mathrm{d}\nu^*}{\mathrm{d}\nu_0} = \frac{Z_0}{Z_*}\exp\Big(\frac{\kappa}{2}m^{*2}\Big), \qquad \mathrm{KL}(\nu^*\|\nu_0) = \frac{\kappa}{2}\|m^*\|^2_{L^2(\nu^*)} - \log\frac{Z_*}{Z_0}, \qquad 1 \le \frac{Z_*}{Z_0} \le \exp\Big(\frac{\kappa}{2}\|u_0^\dagger\|^2\Big).$$

(By Theorem 226 and $s^* = -\lambda m^*$*,* $\frac{\mathrm{d}\nu^*}{\mathrm{d}\nu_0} = \frac{Z_0}{Z_*}\exp(\frac{s^{*2}}{2\lambda\beta}) = \frac{Z_0}{Z_*}\exp(\frac{\lambda m^{*2}}{2\beta})$*; integrating the logarithm against* ν^* *gives the formula for* $\mathrm{KL}(\nu^*\|\nu_0)$*;* $\tilde Z \ge 1$ *since the integrand is* ≥ 1*; and* $\mathrm{KL} \ge 0$ *with Theorem 396 give* $\log\tilde Z \le \frac{\kappa}{2}\|m^*\|^2_{L^2(\nu^*)} \le \frac{\kappa}{2}\|u_0^\dagger\|^2$*.)*

Proof. The exponent identity ‘κ m*²/2 = s*²/(2λβ) = −W‘.
Step 1: the density.
Step 2: the relative entropy, by the Gibbs variational identity with ‘β = 1‘.
Step 3: the bounds on ‘Z̃‘. □

6.3 Convergence to the canonical ridgelet transform

Theorem 402. *For every* $g\colon Z \to \mathbb{R}$*,* $\int g\,\mathrm{d}\nu^* = \int w\,g\,\mathrm{d}\nu_0$*, since* $\nu^* = w\,\nu_0$*.*

Proof. □

Theorem 403. $w^{-1} = \tilde Z\exp(-\frac{\kappa}{2}m^{*2}) \le \tilde Z$*, so* $\|m^*\|^2_{L^2(\nu_0)} = \int m^{*2}w^{-1}\,\mathrm{d}\nu^* \le \tilde Z\|m^*\|^2_{L^2(\nu^*)} \le \exp(\frac{\kappa}{2}\|u_0^\dagger\|^2)\|u_0^\dagger\|^2$.

Proof. Step 1: ‘m² ≤ Z̃ w m²‘ pointwise, since ‘w ≥ 1/Z̃‘.
Step 2: ‘‖m*‖²_L²(ν*) ≤ ‖u₀†‖²‘ (energy bound) and ‘Z̃ ≤ exp((κ/2)‖u₀†‖²)‘. □

Theorem 404. *For* $T > 0$ *put* $\epsilon(T) := (1 - \tilde Z^{-1}) + (e^{\kappa T^2/2} - 1)$*. Where* $|m^*| \le T$ *one has* $\tilde Z^{-1} \le w \le e^{\kappa T^2/2}$*, hence* $|w-1| \le \epsilon(T)$*; where* $|m^*| > T$ *one has* $|w-1| \le w+1$ *and* $|m^*| \le m^{*2}/T$*. Hence pointwise* $|m^*||w-1| \le \epsilon(T)|m^*| + \frac{m^{*2}}{T}(w+1)$ *and*

$$\int |m^*||w-1|\,\mathrm{d}\nu_0 \le \epsilon(T)\|m^*\|_{L^1(\nu_0)} + \frac{1}{T}\big(\|m^*\|^2_{L^2(\nu^*)} + \|m^*\|^2_{L^2(\nu_0)}\big).$$

Proof. Step 1: the pointwise inequality.
Step 2: integrate. □

Theorem 405. *$F_{\rho^*} = S_{\nu^*}m^* = \int m^*\varphi_z\, w\, \mathrm{d}\nu_0$ and $|\varphi_z| \leq 1$ give $\sup_x |F_{\rho^*}(x) - (S_{\nu_0}m^*)(x)| \leq \int |m^*||w-1|\,\mathrm{d}\nu_0$.*

Proof. □

Theorem 406. *$\Pi\rho^* = m^*\nu^* = h\,\nu_0$ with $h := m^*w = \frac{\mathrm{d}\Pi\rho^*}{\mathrm{d}\nu_0} \in L^1(\nu_0)$ (Theorem 229 and $\nu^* = w\nu_0$).*

Proof. □

Theorem 407. *For $u \in L^2(\nu_0)$, $\Pi\rho^* - u\nu_0 = (m^*w - u)\nu_0$, so $\|\Pi\rho^* - u\nu_0\|_{\mathcal{M}} = \|m^*w - u\|_{L^1(\nu_0)} \leq \int |m^*||w-1|\,\mathrm{d}\nu_0 + \|m^* - u\|_{L^1(\nu_0)} \leq \int |m^*||w-1|\,\mathrm{d}\nu_0 + \|m^* - u\|_{L^2(\nu_0)}$.*

Proof. Step 1: the total mass is the ‘$L^1(\nu_0)$‘ norm of the density ‘w m* − u‘.
Step 2: ‘|w m − u| ≤ |m| |w − 1| + |m − u|‘ and ‘$\int$ |m − u| dν_0 ≤ ‖m − u‖_L²(ν_0)‘. □

Theorem 408. *For g measurable with $|g| \leq C$ and $u \in L^2(\nu_0)$, $\left|\int g\ \mathrm{d}\Pi\rho^* - \int g\,u\ \mathrm{d}\nu_0\right| = \left|\int g\,(m^*w - u)\,\mathrm{d}\nu_0\right| \leq C\,\|\Pi\rho^* - u\nu_0\|_{\mathcal{M}}$.*

Proof. □

Theorem 409. *$\mathrm{KL}(\nu_n^*\|\nu_0) \leq \frac{\kappa_n}{2}\|u_0^\dagger\|^2 \to 0$ (Theorem 398 with (λ_n, β_n)).*

Proof. □

Theorem 410. *$\|\nu_n^* - \nu_0\|_{\mathrm{TV}} \leq \sqrt{\mathrm{KL}(\nu_n^*\|\nu_0)/2} \to 0$ (Pinsker's inequality, Lemma 850).*

Proof. □

Theorem 411. *$\nu_n^* \rightharpoonup \nu_0$: for every $g \in C_b(Z)$, $|\int g\,\mathrm{d}\nu_n^* - \int g\,\mathrm{d}\nu_0| \leq 2\|g\|_\infty\|\nu_n^* - \nu_0\|_{\mathrm{TV}} \to 0$.*

Proof. □

Theorem 412. *$\|F_{\rho_n^*} - f\|^2_{L^2(P_X)} \leq \lambda_n\|u_0^\dagger\|^2 \to 0$ (Theorem 398).*

Proof. □

Theorem 413. *$\|m_n^*\|_{L^2(\nu_n^*)} \leq \|u_0^\dagger\|_{L^2(\nu_0)}$ for all n (Theorem 396).*

Proof. □

Theorem 414. *$1 \leq \frac{Z_{*,n}}{Z_0} = \tilde{Z}_n \leq \exp(\frac{\kappa_n}{2}\|u_0^\dagger\|^2) \to 1$ (Theorem 401).*

Proof. □

Theorem 415. *$\|m_n^*\|^2_{L^2(\nu_0)} \leq \tilde{Z}_n\|m_n^*\|^2_{L^2(\nu_n^*)} \leq \exp(\frac{\kappa_n}{2}\|u_0^\dagger\|^2)\|u_0^\dagger\|^2$. In particular (m_n^*) is bounded in $L^2(\nu_0)$ and $\limsup_n \|m_n^*\|_{L^2(\nu_0)} \leq \|u_0^\dagger\|$.*

Proof. □

Theorem 416. *$\int_Z |m_n^*||w_n - 1|\,\mathrm{d}\nu_0 \to 0$; in particular $\sup_x |F_{\rho_n^*}(x) - (S_{\nu_0}m_n^*)(x)| \leq \int |m_n^*||w_n - 1|\,\mathrm{d}\nu_0 \to 0$. (Truncation, Lemma 404: for every $T > 0$, $\int |m_n^*||w_n - 1|\,\mathrm{d}\nu_0 \leq \epsilon_n(T)\sqrt{\tilde{Z}_n}\|u_0^\dagger\| + \frac{1}{T}(1 + \tilde{Z}_n)\|u_0^\dagger\|^2$ with $\epsilon_n(T) \to 0$ and $\tilde{Z}_n \to 1$, so $\limsup_n \int |m_n^*||w_n - 1|\,\mathrm{d}\nu_0 \leq 2\|u_0^\dagger\|^2/T$; let $T \to \infty$.)*

Proof. Step 1: the truncation bound at level ‘T‘, uniformly in ‘n‘.
Step 2: the bound tends to ‘2‖u₀†‖²/T‘ as ‘n → ∞‘.
Step 3: conclude, choosing ‘T‘ with ‘2‖u₀†‖²/T < ε‘. ☐

Theorem 417. $\|S_{\nu_0} m_n^* - f\|_{L^2(P_X)} \le \|S_{\nu_0} m_n^* - F_{\rho_n^*}\| + \|F_{\rho_n^*} - f\| \le \int |m_n^*||w_n - 1| \,\mathrm{d}\nu_0 + \sqrt{\lambda_n}\|u_0^\dagger\| \to 0$ *(Lemma 416 and Theorem 412).*

Proof. ☐

Theorem 418. $\langle m_n^*, u_0^\dagger\rangle_{L^2(\nu_0)} \to \|u_0^\dagger\|^2$. *(For $\varepsilon > 0$ pick $g \in L^2(P_X)$ with $\|u_0^\dagger - S_{\nu_0}^* g\| \le \varepsilon$, possible since $u_0^\dagger \in (\ker S_{\nu_0})^\perp = \overline{\operatorname{ran} S_{\nu_0}^*}$; then $\langle m_n^*, S_{\nu_0}^* g\rangle = \langle S_{\nu_0} m_n^*, g\rangle \to \langle f, g\rangle = \langle u_0^\dagger, S_{\nu_0}^* g\rangle$ by Theorem 417, and the remaining terms are bounded by $(\sup_n \|m_n^*\|_{L^2(\nu_0)} + \|u_0^\dagger\|)\varepsilon$ by Lemma 415.)*

Proof. Step 1: ‘‖m_n^*‖_L²(ν₀) ≤ 2‖u₀†‖‘ eventually (‘Z̃_n ≤ 4‘ eventually).
Step 2: ‘⟪m_n^*, S* g⟫ → ⟪u₀†, S* g⟫‘ for every ‘g ∈ L²(P)‘.
Step 3: the ‘ε/3‘ argument. ☐

Theorem 419. $m_n^* \to u_0^\dagger$ *in* $L^2(\nu_0)$*, i.e.* $\|m_n^* - u_0^\dagger\|_{L^2(\nu_0)} \to 0$. *(In Lean, without compactness:* $\|m_n^* - u_0^\dagger\|^2 = \|m_n^*\|^2 - 2\langle m_n^*, u_0^\dagger\rangle + \|u_0^\dagger\|^2 \le \tilde{Z}_n\|u_0^\dagger\|^2 - 2\langle m_n^*, u_0^\dagger\rangle + \|u_0^\dagger\|^2 \to 0$ *by Lemmas 415 and 418.)*

Proof. ☐

Theorem 420. $\|m_n^*\|_{L^2(\nu_0)} \to \|u_0^\dagger\|$, $\|m_n^*\|_{L^2(\nu_n^*)} \to \|u_0^\dagger\|$ *and* $\|S_{\nu_0} m_n^* - f\|_{L^2(P_X)} \to 0$. *(The first from (a); for the second,* $\tilde{Z}_n^{-1}\|m_n^*\|^2_{L^2(\nu_0)} \le \|m_n^*\|^2_{L^2(\nu_n^*)} \le \|u_0^\dagger\|^2$ *with* $\tilde{Z}_n \to 1$*; the third is* (**??**)*.)*

Proof. ☐

Theorem 421. *With* $h_n := m_n^* w_n = \frac{\mathrm{d}\gamma_n}{\mathrm{d}\nu_0} \in L^1(\nu_0)$, $\|h_n - u_0^\dagger\|_{L^1(\nu_0)} = \|\gamma_n - \gamma_\infty\|_{\mathcal{M}} \to 0$*:* $\|h_n - u_0^\dagger\|_{L^1(\nu_0)} \le \int |m_n^*||w_n - 1| \,\mathrm{d}\nu_0 + \|m_n^* - u_0^\dagger\|_{L^2(\nu_0)} \to 0$ *by Lemma 416 and (a).*

Proof. ☐

Theorem 422. *For every bounded measurable g,* $|\int g \,\mathrm{d}\Pi\rho_n^* - \int g\, u_0^\dagger \,\mathrm{d}\nu_0| \le \|g\|_\infty \|\Pi\rho_n^* - u_0^\dagger\nu_0\|_{\mathcal{M}} \to 0$.

Proof. ☐

Theorem 423. $\gamma_n = \Pi\rho_n^* \rightharpoonup \gamma_\infty = u_0^\dagger\nu_0$*, and* $\|m_n^*\|_{L^2(\nu_n^*)} \to \|u_0^\dagger\|_{L^2(\nu_0)}$*: the learned coefficient measure converges to the canonical (minimum-norm) ridgelet transform with respect to the reference measure* ν_0*. (In Lean this is deduced from Theorem 421: convergence in total mass implies weak convergence.)*

Proof. ☐

Theorem 424. $L(\rho_n^*) = o(\lambda_n)$, $\mathrm{KL}(\nu_n^*\|\nu_0) + \mathrm{KL}(\nu_0\|\nu_n^*) = o(\kappa_n)$ *and* $\|u_0^\dagger - m_n^*\|_{L^2(\nu_0)} \to 0$. *(The Pythagorean identity* (**??**) *with* $u_0 = u_0^\dagger$ *and* (λ_n, β_n) *gives* $\frac{4}{\lambda_n}L(\rho_n^*) + \|u_0^\dagger - m_n^*\|^2_{L^2(\nu_0)} + \frac{2}{\kappa_n}[\mathrm{KL}(\nu_n^*\|\nu_0) + \mathrm{KL}(\nu_0\|\nu_n^*)] = \|u_0^\dagger\|^2 - \|m_n^*\|^2_{L^2(\nu_n^*)} \to 0$ *by Theorem 420; the three terms on the left are nonnegative.)*

Proof. The defect ‘‖u₀†‖² − ‖m_n^*‖²_L²(ν_n^*) → 0‘. ☐

6.4 Order of the limits

Theorem 425. *Assume (R). Let g be bounded measurable (e.g. $g \in C_b(Z)$) and $\varepsilon > 0$. There are $\lambda_0 = \lambda_0(\varepsilon, g) > 0$ and $\kappa_0 = \kappa_0(\varepsilon, g) > 0$ such that every (λ, β) with $\lambda \le \lambda_0$ and $\kappa = \lambda/\beta \le \kappa_0$ satisfies $\left|\int g \,\mathrm{d}\Pi\rho^*_{\lambda,\beta} - \int g\, u_0^\dagger \,\mathrm{d}\nu_0\right| \le \varepsilon$ (the paper writes $\varepsilon/3$). (If the claim were false, there would be minimizers ρ^*_k for (λ_k, β_k) with $\lambda_k, \kappa_k \le 1/(k+1)$ violating the bound; this contradicts $\Pi\rho^*_k \rightharpoonup u_0^\dagger \nu_0$ (Theorem 422) along that schedule.)*

Proof. Step 1: a violating minimizer for ‘λ₀ = κ₀ = 1/(k+1)‘, for every ‘k‘.
Step 2: the violating family is a schedule in the sense of ‘def:m4-schedule‘.
Step 3: along the schedule the integrals converge, contradicting the violation. □

Theorem 426. *In the setting of Theorem 328 with the complexity bound $\mathfrak{R}_N(\Phi) \le C\sqrt{(m+1)/N}$, for fixed (λ, β), a bounded measurable g, $\varepsilon > 0$ and $\delta \in (0,1)$ there is $N_1 = N_1(\varepsilon, \delta, g; \lambda, \beta)$ such that for $N \ge N_1$, with probability at least $1-\delta$, $\left|\int g \,\mathrm{d}\Pi\rho^*_N - \int g \,\mathrm{d}\Pi\rho^*_{\lambda,\beta}\right| \le 2\|g\|_\infty \|\Pi\rho^*_N - \Pi\rho^*_{\lambda,\beta}\|_{\mathrm{TV}} \le 2\|g\|_\infty r_N \le \varepsilon$ (the paper writes $\varepsilon/3$). (Corollary 340 on the event E'_δ of Lemma 339, the rate Corollary 341, and $r_N \to 0$.)*

Proof. Step 1: ‘2 ‖g‖_∞ r_N ≤ ε‘ for ‘N ≥ N₀‘.
Step 2: on the event ‘E'_δ‘ (probability ‘≥ 1 − δ‘), the three bounds. □

Theorem 427. *Assume the setting of Theorem 328 ((A1)–(A5), i.i.d. sample, $|Y| \le Y_{\max}$, $f = \mathbb{E}[Y \mid X]$), (R), the convention (**??**) and $\mathfrak{R}_N(\Phi) \le C\sqrt{(m+1)/N}$. Let g be bounded measurable, $\varepsilon > 0$ and $\delta \in (0,1)$. Then there are $\lambda_0, \kappa_0 > 0$ such that for every (λ, β) with $\lambda \le \lambda_0$, $\lambda/\beta \le \kappa_0$ there is $N_1 = N_1(\varepsilon, \delta, g; \lambda, \beta)$ such that for $N \ge N_1$, with probability at least $1-\delta$,*

$$\left|\int g \,\mathrm{d}\Pi\rho^*_N - \int g\, u_0^\dagger \,\mathrm{d}\nu_0\right| \le \underbrace{2\|g\|_\infty \|\Pi\rho^*_N - \Pi\rho^*\|_{\mathrm{TV}}}_{\le \varepsilon/3} + \underbrace{\left|\int g \,\mathrm{d}(\Pi\rho^* - u_0^\dagger \nu_0)\right|}_{\le \varepsilon/3} \le \frac{2\varepsilon}{3}.$$

*(Stages (1) and (2) of Theorem **??** and the triangle inequality; the third stage, the reachability $e^{(g)}_{M,t}$, is the subject of the dynamics.)*

Proof. Stage (1) at ‘(λ, β)‘.
Stage (2) at ‘(λ, β)‘. □

6.5 The zero-temperature limit: total-variation regularization

Theorem 428. *If $\gamma = h\nu$ with $h \in L^1(\nu)$ then $(S\gamma)(x) = \int_Z h(z)\varphi_z(x)\,\nu(\mathrm{d}z) = (S_\nu h)(x)$ for every x.*

Proof. □

Theorem 429. *If $\gamma = h\nu$ with $h \in L^1(\nu)$ then $S\gamma = S_\nu h$.*

Proof. □

Theorem 430. *For every finite signed measure γ on Z and every $x \in \mathcal{X}$, $|(S\gamma)(x)| \le \|\gamma\|_{\mathcal{M}}$, since $|\varphi_z(x)| \le 1$.*

Proof. □

Theorem 431. *For every finite signed measure γ on Z the synthesis $S\gamma\colon \mathcal{X} \to \mathbb{R}$ is measurable.*

Proof. □

Theorem 432. $S(\gamma_1 + \gamma_2) = S\gamma_1 + S\gamma_2$ *pointwise.*

Proof. □

Theorem 433. $S(c\gamma) = c\, S\gamma$ *pointwise, for* $c \in \mathbb{R}$.

Proof. □

Theorem 434. *If* $\int |a| \,\mathrm{d}\rho < \infty$ *then* $F_\rho(x) = \int_Z \varphi_z(x)\, \Pi\rho(\,\mathrm{d}z) = (S\,\Pi\rho)(x)$ *for every* $x \in \mathcal{X}$.

Proof. □

Theorem 435. *If* $\int |a| \,\mathrm{d}\rho < \infty$ *then* $\|\Pi\rho\|_{\mathcal{M}} \le \int |a| \,\mathrm{d}\rho$, *so that* $\sup_x |F_\rho(x)| \le \|\Pi\rho\|_{\mathcal{M}} \le \int |a| \,\mathrm{d}\rho$.

Proof. Step 1: ‘Πρ = (snd)_*(a⁺ ρ) − (snd)_*(a⁻ ρ)‘, so ‘‖Πρ‖_ℳ ≤ ∫ a⁺ dρ + ∫ a⁻ dρ‘.
Step 2: ‘∫ a⁺ dρ + ∫ a⁻ dρ = ∫ |a| dρ‘. □

Theorem 436. *If* $\int |a| \,\mathrm{d}\rho < \infty$ *then* $L(\rho) = L(\Pi\rho)$, *where* $L(\gamma) = \frac{1}{2}\|S\gamma - f\|^2_{L^2(P_X)}$.

Proof. □

Theorem 437. *Let* $\rho \in \mathcal{P}_2$ *with* $\Pi\rho = \gamma$. *Then* $\|\gamma\|_{\mathcal{M}} \le \int |a| \,\mathrm{d}\rho \le \left(\int a^2 \,\mathrm{d}\rho\right)^{1/2}$, *i.e.* $\|\gamma\|^2_{\mathcal{M}} \le \int a^2 \,\mathrm{d}\rho$.

Proof. □

Theorem 438. *The measure* $\bar\rho$ *of* (159) *is a probability measure in* $\mathcal{P}_2$ *with* $\Pi\bar\rho = \gamma$ *and* $\int a^2 \,\mathrm{d}\bar\rho = \|\gamma\|^2_{\mathcal{M}}$; *for* $\gamma = 0$ *the same holds for* $\delta_0 \otimes \nu$.

Proof. □

Theorem 439. *Let* γ *be a finite signed measure on* Z *(and* $\nu \in \mathcal{P}(Z)$ *arbitrary, used only when* $\gamma = 0$*). Then* $\inf\{\int a^2 \,\mathrm{d}\rho : \rho \in \mathcal{P}_2,\ \Pi\rho = \gamma\} = \|\gamma\|^2_{\mathcal{M}}$ *and the infimum is attained, at* $\bar\rho$ *of* (159).

Proof. □

Theorem 440. $\inf\{\int a^2 \,\mathrm{d}\rho : \rho \in \mathcal{P}_2,\ \Pi\rho = \gamma\} = \|\gamma\|^2_{\mathcal{M}}$.

Proof. □

Theorem 441. *For* $\rho \in \mathcal{P}_2$ *and* $\lambda \ge 0$, $L(\rho) + \frac{\lambda}{2}\int a^2 \,\mathrm{d}\rho \ge L(\Pi\rho) + \frac{\lambda}{2}\|\Pi\rho\|^2_{\mathcal{M}}$.

Proof. □

Theorem 442. *For every* $\gamma \in \mathcal{M}(Z)$ *the measure* $\bar\rho$ *of* (159) *satisfies* $L(\bar\rho) + \frac{\lambda}{2}\int a^2 \,\mathrm{d}\bar\rho = L(\gamma) + \frac{\lambda}{2}\|\gamma\|^2_{\mathcal{M}}$.

Proof. □

Theorem 443. *Let* $\lambda \ge 0$, P *a measure on* $\mathcal{X}$, $f\colon \mathcal{X} \to \mathbb{R}$ *and* $\nu \in \mathcal{P}(Z)$ *(so that* $\mathcal{P}_2 \ne \emptyset$*). Then* $\inf_{\rho\in\mathcal{P}_2}\{L(\rho) + \frac{\lambda}{2}\int a^2 \,\mathrm{d}\rho\} = \inf_{\gamma\in\mathcal{M}(Z)}\{L(\gamma) + \frac{\lambda}{2}\|\gamma\|^2_{\mathcal{M}}\}$, *both sides being infima of nonnegative sets.*

Proof. ‘≤‘: for every ‘γ‘, ‘ρ̄ ∈ $\mathcal{P}_2$‘ has ‘$\mathcal{E}_0$(ρ̄) = ℰ_TV(γ)‘.
‘≥‘: for every ‘ρ ∈ $\mathcal{P}_2$‘, ‘ℰ_TV(Πρ) ≤ $\mathcal{E}_0$(ρ)‘. □

Theorem 444. *If $\rho \in \mathcal{P}_2$ minimizes $L(\rho) + \frac{\lambda}{2}\int a^2 \,\mathrm{d}\rho$ over $\mathcal{P}_2$ then $\Pi\rho$ minimizes $L(\gamma) + \frac{\lambda}{2}\|\gamma\|_{\mathcal{M}}^2$ over $\mathcal{M}(Z)$.*

Proof. □

Theorem 445. *If γ minimizes $L(\gamma) + \frac{\lambda}{2}\|\gamma\|_{\mathcal{M}}^2$ over $\mathcal{M}(Z)$ then $\bar{\rho}$ of (159) lies in $\mathcal{P}_2$ and minimizes $L(\rho) + \frac{\lambda}{2}\int a^2 \,\mathrm{d}\rho$ over $\mathcal{P}_2$.*

Proof. □

Theorem 446. *For $f \in L^2(P_X)$ the functional $\gamma \mapsto L(\gamma) = \frac{1}{2}\|S\gamma - f\|_{L^2(P_X)}^2$ is convex on $\mathcal{M}(Z)$.*

Proof. Step 1: '$(S\gamma - f)^2 \in L^1(P)$' for every '$\gamma$', since '$S\gamma$' is bounded and '$f \in L^2$'.
Step 2: the pointwise convexity inequality '$(a\, s_1 + b\, s_2 - f)^2 \le a\,(s_1 - f)^2 + b\,(s_2 - f)^2$', integrated.
□

Theorem 447. *$\gamma \mapsto \|\gamma\|_{\mathcal{M}}^2$ is convex on $\mathcal{M}(Z)$, since $\|\cdot\|_{\mathcal{M}}$ is a norm.*

Proof. □

Theorem 448. *For $f \in L^2(P_X)$ and $\lambda \ge 0$ the functional $\gamma \mapsto L(\gamma) + \frac{\lambda}{2}\|\gamma\|_{\mathcal{M}}^2$ is convex on $\mathcal{M}(Z)$.*

Proof. □

Chapter 7

Convergence rates for the regularization limit

Section 10 of the paper: the learned-base-measure problem rewritten as a weighted Tikhonov problem over the reference measure, the pointwise deviation of the weight, the sup-norm bound of order $\lambda^{-1/2}$, the Tikhonov rate under a source condition (of order $1/2$ or 1), the rate theorem, the rates under the uniform sup-norm bound and in the high-temperature regime, and the finite-dimensional case.

7.1 Rewriting as a weighted Tikhonov problem

Theorem 449. *For $\lambda > 0$, $u_\lambda = S^*_{\nu_0}(K_{\nu_0} + \lambda)^{-1} f$ satisfies $(T_0 + \lambda)u_\lambda = S^*_{\nu_0} f$ (Lemma **??**(4)).*

Proof. □

Theorem 450. *If $f = K_{\nu_0} g = S_{\nu_0}(S^*_{\nu_0} g)$ with $g \in L^2(P_X)$, then (R) holds and $u^\dagger = S^*_{\nu_0} g$, since $S^*_{\nu_0} g \in \operatorname{ran} S^*_{\nu_0} \subseteq (\ker S_{\nu_0})^\perp$ (Lemma 384).*

Proof. □

Theorem 451. *If $f = S_{\nu_0} T_0 g_0$ with $g_0 \in L^2(\nu_0)$, then (R) holds and $u^\dagger = T_0 g_0$, since $T_0 g_0 = S^*_{\nu_0}(S_{\nu_0} g_0) \in (\ker S_{\nu_0})^\perp$.*

Proof. □

Theorem 452. *(SC$_a$) implies (R): $f = S_{\nu_0} u^\dagger$.*

Proof. □

Theorem 453. *Assume (A1), (A3) and (R). Then $u_\lambda - u^\dagger = -\lambda(T_0 + \lambda)^{-1} u^\dagger$, i.e. $(T_0+\lambda)(u_\lambda - u^\dagger) = -\lambda u^\dagger$. ($f = S_{\nu_0} u^\dagger$ gives $S^*_{\nu_0} f = T_0 u^\dagger$, so $(T_0 + \lambda)u_\lambda - (T_0 + \lambda)u^\dagger = T_0 u^\dagger - (T_0 + \lambda)u^\dagger = -\lambda u^\dagger$.)*

Proof. □

Theorem 454. *Under (SC$_{1/2}$), $f = K_{\nu_0} g$ with $g \in L^2(P_X)$: $\|u_\lambda - u^\dagger\|_{L^2(\nu_0)} \le \frac{\sqrt{\lambda}}{2}\|g\| \le \sqrt{\lambda}\|g\|$ and $\|S_{\nu_0}(u_\lambda - u^\dagger)\|_{L^2(P_X)} \le \lambda\|g\|$. (With $y := (T_0 + \lambda)^{-1} S^*_{\nu_0} g$ one has $u_\lambda - u^\dagger = -\lambda y$, and the normal equation $(T_0 + \lambda)y = S^*_{\nu_0} g$ gives $\|S_{\nu_0} y\|^2 + \lambda\|y\|^2 \le \|g\|\,\|S_{\nu_0} y\|$, hence $\lambda\|y\|^2 \le \|g\|^2/4$ and $\|S_{\nu_0} y\| \le \|g\|$.)*

Proof. ‘y := −λ⁻¹ d‘ solves the normal equation ‘(T₀ + λ) y = S_ν₀* g‘.
‘‖d‖² = λ²‖y‖² ≤ λ‖g‖²/4‘. ☐

Theorem 455. *Under (SC$_1$), $u^\dagger = T_0 g_0$ with $g_0 \in L^2(\nu_0)$: $\|u_\lambda - u^\dagger\|_{L^2(\nu_0)} \le \lambda\|g_0\|$ and $\|S_{\nu_0}(u_\lambda - u^\dagger)\|_{L^2(P_X)} \le \lambda\|g_0\|$. ($u_\lambda - u^\dagger = -\lambda(T_0+\lambda)^{-1}T_0 g_0$ and $\|T_0(T_0+\lambda)^{-1}\| \le 1$; the residual bound follows from $\|S_{\nu_0}\| \le 1$.)*

Proof. ☐

Theorem 456. *Assume (A1), (A3) and (SC$_a$) (hence (R)). Then*

$$\|u_\lambda - u^\dagger\|_{L^2(\nu_0)} \le \lambda^{\min(a,1)}\|g_0\|_{L^2(\nu_0)}, \qquad \|S_{\nu_0}(u_\lambda - u^\dagger)\|_{L^2(P_X)} \le \lambda^{\min(a+1/2,1)}\|g_0\|_{L^2(\nu_0)}.$$

(For $a \in \{1/2, 1\}$: Lemmas 454 and 455, with $\lambda^{1/2} = \sqrt{\lambda}$ and $\min(a+1/2,1) = 1$.)

Proof. ☐

Theorem 457. *With $v := w\,m^*$: v is bounded, $v \in L^2(\nu_0)$ and $S_{\nu_0}v = S_{\nu^*}m^* = F_{\rho^*}$ (by $|m^*\varphi_z| \le \|m^*\|_\infty$ and Fubini, $S_{\nu^*}m^*(x) = \int m^*(z)\varphi_z(x)w(z)\nu_0(\,\mathrm{d}z) = (S_{\nu_0}v)(x)$, which is F_{ρ^*} by Theorem 231).*

Proof. ☐

Theorem 458. $\Pi\rho^* = m^*\nu^* = v\,\nu_0$ *and* $\|m^*\|^2_{L^2(\nu^*)} = \int m^{*2}w\,\mathrm{d}\nu_0 = \int v^2 w^{-1}\,\mathrm{d}\nu_0$.

Proof. ☐

Theorem 459. (Weighted normal equation.) *With $M_{w^{-1}}$ the multiplication operator by w^{-1}, in $L^2(\nu_0)$ and pointwise, $(T_0 + \lambda M_{w^{-1}})\,v = S^*_{\nu_0}f$, where $w^{-1}v = m^*$. (By Theorem 228, $s^* = S^*r^* = -\lambda m^*$ for every z; since $r^* = F_{\rho^*} - f = S_{\nu_0}v - f$, $S^*r^* = T_0 v - S^*_{\nu_0}f$, hence $T_0 v - S^*_{\nu_0}f = -\lambda m^* = -\lambda w^{-1}v$.)*

Proof. ☐

Theorem 460. (Variational interpretation.) *v is the unique minimizer of the weighted Tikhonov functional $\mathcal{J}_w(u) := \frac{1}{2}\|S_{\nu_0}u - f\|^2_{L^2(P_X)} + \frac{\lambda}{2}\int_Z u^2 w^{-1}\,\mathrm{d}\nu_0$ ($u \in L^2(\nu_0)$): for every u, $\mathcal{J}_w(u) = \mathcal{J}_w(v) + \frac{1}{2}\|S_{\nu_0}(u-v)\|^2 + \frac{\lambda}{2}\int (u-v)^2 w^{-1}\,\mathrm{d}\nu_0$, since the cross terms vanish by the weighted normal equation.*

Proof. Integrability of the weighted squares and of the cross term.
Step 1: the weighted integral splits, since ‘v w⁻¹ = m‘.
Step 2: the cross term is ‘⟪d, T₀ v − S* f + λ m⟫ = 0‘ by the normal equation.
Step 3: expand the square of the residual. ☐

Theorem 461. (Difference from u_λ.) *$(T_0+\lambda)(v - u_\lambda) = \lambda(1 - w^{-1})v = \lambda(w-1)m^*$ and $\|v - u_\lambda\|_{L^2(\nu_0)} \le \|(w-1)m^*\|_{L^2(\nu_0)} = \Xi$. (Subtract $(T_0+\lambda)u_\lambda = S^*_{\nu_0}f$ from the weighted normal equation, and use $\|\lambda(T_0+\lambda)^{-1}\| \le 1$.)*

Proof. ☐

Theorem 462. *Assume the hypotheses of Lemma 459 and (R). Then $w = q^{-1}e^{\zeta}$ with $q = Z_*/Z_0 = \tilde{Z}$, and pointwise*

$$e^{-\zeta_0} \le q^{-1} \le w(z) \le e^{\zeta(z)}, \qquad |w(z)-1| \le \zeta(z)e^{\zeta(z)} + \zeta_0, \qquad |1-w(z)^{-1}| \le \zeta(z) + e^{\zeta_0} - 1 \le \zeta(z) + \zeta_0 e^{\zeta_0}.$$

($1 \le q \le e^{\zeta_0}$; $e^{\zeta} - 1 \le \zeta e^{\zeta}$; $1 - q^{-1} \le \log q \le \zeta_0$; $1 - e^{-\zeta} \le \zeta$.)

Proof. ‘e^−$\zeta_0 \le 1/\tilde{Z}$‘ since ‘$\tilde{Z} \le$ e^ζ_0‘.

‘$1 - 1/\tilde{Z} \le \log \tilde{Z} \le \zeta_0$‘.

‘$w^{-1} = \tilde{Z}$ e^−ζ‘. ☐

Theorem 463. *Pointwise, $|(w-1)m^*|(z) \le \frac{\kappa}{2}|m^*(z)|^3 e^{\kappa m^*(z)^2/2} + \frac{\kappa}{2}\|u^\dagger\|^2|m^*(z)|$ (insert $\zeta = \frac{\kappa}{2}m^{*2}$, $\zeta_0 = \frac{\kappa}{2}\|u^\dagger\|^2$ into the bound on $|w-1|$ and multiply by $|m^*|$).*

Proof. ☐

Theorem 464. $\Xi \le \frac{\kappa}{2}(\|m^{*3}e^{\kappa m^{*2}/2}\|_{L^2(\nu_0)} + \|u^\dagger\|^2\,\|m^*\|_{L^2(\nu_0)})$ *(triangle inequality in $L^2(\nu_0)$ applied to the pointwise bound of Lemma 463).*

Proof. Step 1: ‘$\Xi \le \|c_1\,\varphi_1 + c_2\,\varphi_2\|$‘ by pointwise domination.

Step 2: the triangle inequality and ‘$\|\,|m^*|\,\| = \|m^*\|$‘. ☐

Theorem 465. $\|m^*\|_{L^2(\nu_0)} \le q^{1/2}\|m^*\|_{L^2(\nu^*)} \le e^{\zeta_0/2}\|u^\dagger\|$ *($\int m^{*2}\,\mathrm{d}\nu_0 = \int m^{*2}w^{-1}\,\mathrm{d}\nu^* \le q\|m^*\|^2_{L^2(\nu^*)}$, $\|m^*\|_{L^2(\nu^*)} \le \|u^\dagger\|$ and $q \le e^{\zeta_0}$).*

Proof. ☐

Theorem 466. *For every $\nu \in \mathcal{P}(Z)$ and $f \in \operatorname{ran} S_\nu$, $\sup_{z\in Z}|R_{\lambda,\nu}f(z)| \le \|S_\nu^\dagger f\|_{L^2(\nu)}/(2\sqrt{\lambda})$; more generally the bound holds with $\|h\|_{L^2(\nu)}$ for every representer h of f. (With $g := (K_\nu+\lambda)^{-1}f$, $|R_{\lambda,\nu}f(z)| \le \|g\|$ and $\|S_\nu^* g\|^2 + \lambda\|g\|^2 = \langle (K_\nu+\lambda)g, g\rangle = \langle h, S_\nu^* g\rangle \le \|h\|\|S_\nu^* g\|$, hence $\lambda\|g\|^2 \le \|h\|^2/4$.)*

Proof. ☐

Theorem 467. *For $m^* = R_{\lambda,\nu^*}f$,*

$$\sup_{z\in Z}|m^*(z)| \le \frac{1}{2\sqrt{\lambda}}\left\|\frac{u^\dagger}{w}\right\|_{L^2(\nu^*)} \le \frac{q^{1/2}\|u^\dagger\|}{2\sqrt{\lambda}} \le \frac{e^{\zeta_0/2}\|u^\dagger\|}{2\sqrt{\lambda}}, \qquad \sup_z \zeta(z) \le \frac{e^{\zeta_0}\|u^\dagger\|^2}{8\beta} =: \zeta_\beta.$$

($h := u^\dagger/w \in L^2(\nu^*)$ with $\|h\|^2_{L^2(\nu^*)} = \int u^{\dagger 2}w^{-1}\,\mathrm{d}\nu_0 \le q\|u^\dagger\|^2$ and $S_{\nu^*}h = S_{\nu_0}u^\dagger = f$; apply Lemma 466.)

Proof. Step 1: ‘h := u†/w ∈ $L^2(\nu^*)$‘, via ‘$\int$ w (u†/w)² dν_0 = $\int$ u†² w^{-1} d$\nu_0 \le \tilde{Z}$ ‖u†‖²‘.

Step 2: ‘S_ν* h = S_ν$_0$ u† = f‘.

Step 3: apply the general bound and insert ‘$\tilde{Z} \le$ e^ζ_0‘.

‘ζ = (κ/2) m*² ≤ (κ/2) $\tilde{Z}$‖u†‖²/(4λ) = $\tilde{Z}$‖u†‖²/(8β) ≤ e^ζ_0‖u†‖²/(8β)‘. ☐

7.2 The rate theorem and its corollaries

Theorem 468. *Assume (A1), (A3), (A4), (A5) and (SC$_a$) (hence (R)). Fix (λ,β) and let $\Xi = \|(w-1)m^*\|_{L^2(\nu_0)}$. Then*

$$\|m^* - u^\dagger\|_{L^2(\nu_0)} \le 2\,\Xi + \lambda^{\min(a,1)}\|g_0\|.$$

(Write $m^* - u^\dagger = (m^* - v) + (v - u_\lambda) + (u_\lambda - u^\dagger)$: the first term is $(1-w)m^*$ of norm Ξ, the second is at most Ξ by Lemma 461, the third is Lemma 456.)

Proof. ☐

Theorem 469. *With* $\Pi\rho^* = m^*\nu^* = v\nu_0$*,*

$$\|\Pi\rho^* - u^\dagger\nu_0\|_{\mathcal{M}} = \|v - u^\dagger\|_{L^1(\nu_0)} \leq 3\,\Xi + \lambda^{\min(a,1)}\|g_0\|.$$

($\|v-u^\dagger\|_{L^1} \leq \|v-m^*\|_{L^1} + \|m^*-u^\dagger\|_{L^1} \leq \|v-m^*\|_{L^2} + \|m^*-u^\dagger\|_{L^2}$*,* ν_0 *being a probability measure.)*

Proof. '∫ |m*| |w − 1| dν0 = ∫ |(w − 1) m*| dν0 ≤ Ξ'. ☐

Theorem 470. $\|F_{\rho^*} - f\|_{L^2(P_X)} \leq \frac{\sqrt{\lambda}}{2}\,\Xi + \lambda^{\min(a+1/2,1)}\|g_0\|$*. (*$F_{\rho^*} - f = S_{\nu_0}(v - u^\dagger) = S_{\nu_0}(v - u_\lambda) + S_{\nu_0}(u_\lambda - u^\dagger)$*;* $\|S_{\nu_0}(v - u_\lambda)\|^2 = \langle T_0 e, e\rangle$ *with* $(T_0 + \lambda)e = \lambda(w-1)m^*$*, and* $4\lambda\langle T_0 e, e\rangle \leq \|(T_0 + \lambda)e\|^2 = \lambda^2\Xi^2$*.)*

Proof. '‖S_ν0 e‖² = ⟪T0 e, e⟫ ≤ ‖(T0 + λ) e‖²/(4λ) = λ Ξ²/4'. ☐

Theorem 471. $\Xi \leq \frac{\kappa}{2}(\|m^{*3}e^{\kappa m^{*2}/2}\|_{L^2(\nu_0)} + \|u^\dagger\|^2\|m^*\|_{L^2(\nu_0)})$ *(Lemma 464).*

Proof. ☐

Theorem 472. *Assume the hypotheses of Theorem 468 and* (S_∞) *with constant* B_∞*, and put* $\zeta_\infty := \frac{\kappa}{2}B_\infty^2$*. Then*

$$\Xi \leq \frac{\kappa}{2}\Big(B_\infty^2 e^{\zeta_\infty} + \|u^\dagger\|^2\Big)e^{\zeta_0/2}\|u^\dagger\|,$$

and for $\kappa \leq \kappa_0$*,* $\Xi \leq C_1\kappa$ *with* C_1 *as in Definition 145. (In the bound of Lemma 462,* $|w-1| \leq \zeta e^{\zeta} + \zeta_0 \leq \zeta_\infty e^{\zeta_\infty} + \zeta_0$*, and* $\|m^*\|_{L^2(\nu_0)} \leq e^{\zeta_0/2}\|u^\dagger\|$ *by Lemma 465.)*

Proof. Monotonicity in 'κ' of the exponential factors. ☐

Theorem 473. *For* $\kappa \leq \kappa_0$*,*

$$\|m^* - u^\dagger\|_{L^2(\nu_0)} \leq 2C_1\kappa + \lambda^{\min(a,1)}\|g_0\|, \qquad \|\Pi\rho^* - u^\dagger\nu_0\|_{\mathcal{M}} \leq 3C_1\kappa + \lambda^{\min(a,1)}\|g_0\|,$$

$$\|F_{\rho^*} - f\| \leq \frac{C_1}{2}\sqrt{\lambda}\,\kappa + \lambda^{\min(a+1/2,1)}\|g_0\|.$$

(Insert $\Xi \leq C_1\kappa$ *into Theorem 468–470.)*

Proof. ☐

Theorem 474. (Hidden marginal.) *Pointwise* $|w-1| \leq \zeta_\infty e^{\zeta_\infty} + \zeta_0$*, so for* $\kappa \leq \kappa_0$*,* $\|w-1\|_{L^\infty(\nu_0)} \leq C_1'\kappa$ *with* C_1' *as in Definition 146,*

$$\|\nu^* - \nu_0\|_{\mathcal{M}} \leq \|w-1\|_{L^1(\nu_0)} \leq C_1'\kappa, \qquad \mathrm{KL}(\nu^*\|\nu_0) \leq \chi^2(\nu^*\|\nu_0) = \int (w-1)^2\,\mathrm{d}\nu_0 \leq (C_1'\kappa)^2.$$

($\|\nu^* - \nu_0\|_{\mathcal{M}} = \int |w-1|\,\mathrm{d}\nu_0$*, and* $\mathrm{KL} \leq \chi^2$ *by Lemma 905.)*

Proof. Step 1: the pointwise bound '|w − 1| ≤ C1' κ'.
Step 2: the total mass and the 'χ²' distance. ☐

Theorem 475. *Assume only the hypotheses of Theorem 468 (*(S_∞) *is not used), and let* $\zeta_\beta = e^{\zeta_0}\|u^\dagger\|^2/(8\beta)$*. Then*

$$\Xi \leq (\zeta_\beta e^{\zeta_\beta} + \zeta_0)e^{\zeta_0/2}\|u^\dagger\|,$$

and for $\beta \geq \beta_0 > 0$*,* $\kappa \leq \kappa_0$*,* $\Xi \leq C_2(\beta^{-1} + \kappa)$ *with* C_2 *as in Definition 147. (Pointwise* $|w-1| \leq \zeta e^{\zeta} + \zeta_0 \leq \zeta_\beta e^{\zeta_\beta} + \zeta_0$ *by Lemma 467, and* $\|m^*\|_{L^2(\nu_0)} \leq e^{\zeta_0/2}\|u^\dagger\|$*.)*

Proof. Step 2: insert ‘ζ0 ≤ ζ00 = ϰ0 U²/2‘ and ‘1/β ≤ 1/β0‘.
‘ζ_β ≤ (1/β) e^ζ00 U²/8‘ and ‘ζ_β ≤ e^ζ00 U²/(8β0)‘.
The first term of ‘C e^ζ0/2 U‘ is at most ‘(1/β) U³ e^ζ00/2 M‘, the second at most ‘ϰ U³ e^ζ00/2 M‘. □

Theorem 476. *For $\beta \geq \beta_0$ and $\kappa \leq \kappa_0$, $\|m^* - u^\dagger\|_{L^2(\nu_0)} \leq 2C_2(\beta^{-1} + \kappa) + \lambda^{\min(a,1)}\|g_0\|$, and likewise for the coefficient measure and the realization as in Corollary 473, with $C_1\kappa$ replaced by $C_2(\beta^{-1} + \kappa)$.*

Proof. □

Theorem 477. *Suppose K_{ν_0} is coercive on $L^2(P_X)$, $\langle K_{\nu_0} h, h\rangle \geq \sigma_0 \|h\|^2$ with $\sigma_0 > 0$ (in the paper: P_X has finitely many atoms and K_{ν_0} is invertible with smallest eigenvalue σ_0). Then (R) holds for every f: $f = S_{\nu_0}(S^*_{\nu_0} K^{-1}_{\nu_0} f)$.*

Proof. □

Theorem 478. *Assume (A1), (A3), (A4), (A5), and that P_X has finitely many atoms $x_1, \ldots, x_n$, so that $L^2(P_X) \cong \mathbb{R}^n$. Suppose K_{ν_0} is invertible on $L^2(P_X)$ with smallest eigenvalue $\sigma_0 > 0$. Then (R) holds automatically, and for $\kappa \leq \kappa_0$,*

$$\sup_z |m^*(z)| \leq \frac{e^{\zeta_0}\|f\|_{L^2(P_X)}}{\sigma_0} \leq \frac{e^{\kappa_0\|u^\dagger\|^2/2}\|f\|}{\sigma_0} =: B_\infty.$$

($K_{\nu^*} = \int \Phi_z \Phi_z^\top w \, \mathrm{d}\nu_0$ and $w \geq e^{-\zeta_0}$ give $K_{\nu^*} \succeq e^{-\zeta_0} K_{\nu_0} \succeq e^{-\zeta_0}\sigma_0 I$, hence $\|(K_{\nu^*} + \lambda)^{-1}\| \leq e^{\zeta_0}/\sigma_0$ and $|m^*(z)| \leq \|(K_{\nu^*} + \lambda)^{-1} f\|$. In Lean the hypothesis is the coercivity $\sigma_0\|h\|^2 \leq \langle K_{\nu_0} h, h\rangle$ for a general P_X.)

Proof. Step 1: ‘⟪K_ν* g, g⟫ = ∫ w (S* g)² dν0 ≥ e^−ζ0 ⟪K_ν0 g, g⟫‘.
Step 2: ‘e^−ζ0 σ0 ‖g‖² ≤ ⟪(K_ν* + λ) g, g⟫ = ⟪f, g⟫ ≤ ‖f‖ ‖g‖‘.
Step 3: ‘|m*(z)| = |(S* g)(z)| ≤ ‖g‖‘. □

7.3 Spectral powers of T_0 and the source condition of general order

Theorem 479. *For $a > 0$, $\operatorname{ran} T_0^a \subseteq (\ker T_0)^\perp = (\ker S_{\nu_0})^\perp$ (Lemma 792; $\ker T_0 = \ker S_{\nu_0}$ since $\langle T_0 u, u\rangle = \|S_{\nu_0} u\|^2$).*

Proof. □

Theorem 480. *If $f = S_{\nu_0} T_0^a g_0$ with $a > 0$, then (R) holds and $u^\dagger = T_0^a g_0$, since $T_0^a g_0 \in (\ker S_{\nu_0})^\perp$ (Lemma 384).*

Proof. □

Theorem 481. *(SC$_a$) implies (R): $f = S_{\nu_0} u^\dagger$.*

Proof. □

Theorem 482. *The source condition of order 1 of Definition 139, $f = S_{\nu_0} T_0 g_0$ with $\|g_0\| \leq G$, is (SC$_1$) of Definition 165, since $T_0^1 = T_0$.*

Proof. □

Theorem 483. *The source condition of order* $1/2$ *of Definition 139,* $f = K_{\nu_0} g$ *with* $\|g\| \le G$*, implies (SC*$_{1/2}$*) of Definition 165 with the same bound:* $\operatorname{ran} S^*_{\nu_0} \subseteq \operatorname{ran} T_0^{1/2}$ *with* $S^*_{\nu_0} g = T_0^{1/2} g_0$*,* $\|g_0\| \le \|g\|$ *(Lemma 797).*

Proof. □

Theorem 484. *Assume (A1), (A3) and (SC*$_a$*) with* $a > 0$ *(hence (R)). Then*

$$\|u_\lambda - u^\dagger\|_{L^2(\nu_0)} \le \lambda^{\min(a,1)} \|g_0\|_{L^2(\nu_0)}, \qquad \|S_{\nu_0}(u_\lambda - u^\dagger)\|_{L^2(P_X)} \le \lambda^{\min(a+1/2,1)} \|g_0\|_{L^2(\nu_0)}.$$

($u_\lambda - u^\dagger = -\lambda(T_0 + \lambda)^{-1} T_0^a g_0$ *(Lemma 453), the eigenvalues of* T_0 *lie in* $[0,1]$*, and Lemma 796 applies;* $\|S_{\nu_0} h\|^2 = \langle T_0 h, h\rangle$*.)*

Proof. ‘(T_0 + λ) d = λ T_0^a (−g_0)‘: the resolvent bounds apply with ‘g = −g_0‘. □

Theorem 485. *Assume (A1), (A3), (A4), (A5) and (SC*$_a$*) with* $a > 0$ *(hence (R)). Fix* (λ, β) *and let* $\Xi = \|(w-1)m^*\|_{L^2(\nu_0)}$*. Then* $\|m^* - u^\dagger\|_{L^2(\nu_0)} \le 2\,\Xi + \lambda^{\min(a,1)} \|g_0\|$ *(Theorem 468 with Lemma 484 in place of Lemma 456).*

Proof. □

Theorem 486. *With* $\Pi\rho^* = m^*\nu^* = v\nu_0$*,* $\|\Pi\rho^* - u^\dagger \nu_0\|_{\mathcal{M}} = \|v - u^\dagger\|_{L^1(\nu_0)} \le 3\,\Xi + \lambda^{\min(a,1)} \|g_0\|$ *(Theorem 469).*

Proof. ‘∫ |m*| |w − 1| dν_0 = ∫ |(w − 1) m*| dν_0 ≤ Ξ‘. □

Theorem 487. $\|F_{\rho^*} - f\|_{L^2(P_X)} \le \frac{\sqrt{\lambda}}{2}\,\Xi + \lambda^{\min(a+1/2,1)} \|g_0\|$ *(Theorem 470).*

Proof. ‘‖S_ν_0 e‖² = ⟪T_0 e, e⟫ ≤ ‖(T_0 + λ) e‖²/(4λ) = λ Ξ²/4‘. □

7.4 Joint schedule of the sample size and the regularization

Theorem 488. *With* $D(\delta/2) \le (B + Y_{\max})^2 [4C_{\mathrm{P}}\sqrt{m+1} + \sqrt{\tfrac{1}{2}\log(4/\delta)}]$ *and* $N^{-1/2} \le N^{-1/4}$ *for* $N \ge 1$*,* $\sqrt{2D(\delta/2)/\sqrt{N}} + D'(\delta)/\sqrt{N} \le (B + Y_{\max})\, C_{\mathrm{stat}}(m,\delta)\, N^{-1/4}$*.*

Proof. Step 1: ‘D(δ/2) ≤ E² S_1‘.
Step 2: ‘√(2D/√N) ≤ E √(2S_1) N^−1/4‘.
Step 3: ‘D’/√N ≤ E S_2 N^−1/4‘. □

Theorem 489. *Assume the setting of Theorem 328, (SC*$_a$*) with* $0 < a \le 1$*, the complexity bound* $\mathfrak{R}_N(\Phi) \le C_{\mathrm{P}}\sqrt{(m+1)/N}$ *and (S*$_\infty$*) for the family* $\lambda \le 1$*,* $\beta = \beta_0$*; let* C_1 *be Definition 145 with* $\kappa_0 = 1/\beta_0$*. With* $\lambda_N := N^{-1/(4(a+2))}$*, for every* $N \ge 1$*, with probability at least* $1 - \delta$*,*

$$\|m_{\rho^*_N} - u^\dagger\|_{L^2(\nu_0)} \le \Big[(2Y_{\max} + \sqrt{2\beta_0/\pi})C_{\mathrm{stat}} + \frac{2C_1}{\beta_0} + \|g_0\|\Big] N^{-\frac{a}{4(a+2)}}.$$

*((**??**): the first term is Corollary 341 and Lemma 488 with* $B + Y_{\max} \le (2Y_{\max} + \sqrt{2\beta_0/\pi})/\lambda$*, the second is Corollary 473; at* $\lambda = \lambda_N$*,* $\lambda_N^{-2} N^{-1/4} = \lambda_N^a$ *and* $\lambda_N \le \lambda_N^a$*.)*

Proof. The balancing: ‘λ_N^−2 N^−1/4 = λ_N^a = N^−a/(4(a+2))‘ and ‘λ_N ≤ λ_N^a‘.

The bias term: ‘cor:rate-sinf‘ with ‘ϰ₀ = 1/β₀‘.

The sample term: ‘cor:m3-convergence‘ (b), (e) and ‘eq:C-stat‘.

The triangle inequality ‘eq:joint-triangle‘. □

Theorem 490. *Assume the setting of Theorem 328, (SC$_a$) with $0 < a \le 1$ and the complexity bound. With $\lambda_N := N^{-1/(4(a+2))}$ and $\beta_N := \lambda_N^{-a}$, for every $N \ge 1$, with probability at least $1-\delta$,*

$$\|m_{\rho_N^*} - u^\dagger\|_{L^2(\nu_0)} \le \Big[(2Y_{\max}+1)C_{\text{stat}} + 4C_2 + \|g_0\|\Big] N^{-\frac{a}{4(a+2)}},$$

where C_2 is Definition 147 with $\beta_0 = \kappa_0 = 1$. ($\beta_N \ge 1$ and $\kappa_N = \lambda_N^{1+a} \le 1$, so Corollary 476 applies with $\beta_0 = \kappa_0 = 1$ and $2C_2(\beta_N^{-1} + \kappa_N) \le 4C_2\lambda_N^a$; moreover $B + Y_{\max} \le (2Y_{\max}+1)/\lambda$ for $\lambda \le 1$, $a \le 1$.)

Proof. The balancing: ‘λ_N^−2 N^−1/4 = λ_N^a = N^−a/(4(a+2))‘.

The high-temperature schedule: ‘β_N ≥ 1‘, ‘ϰ_N ≤ 1‘, ‘1/β_N = λ_N^a‘, ‘ϰ_N = λ_N^1+a ≤ λ_N^a‘.

The bias term: ‘cor:rate-high-temperature‘ with ‘β₀ = ϰ₀ = 1‘.

The sample term: ‘cor:m3-convergence‘ (b), (e) and ‘eq:C-stat‘.

The triangle inequality ‘eq:joint-triangle‘. □

Theorem 491. *Assume the setting of Theorem 328, (SC$_a$) with $0 < a \le 1$, the complexity bounds of Lemmas 314 and 315, and let $D' = D'(\delta)$, $D'' = D''(\delta)$ be as in (**??**). Put $\lambda_N := N^{-1/(2(a+3))}$ and $\beta_N := \lambda_N^{-a}$. Then for $N^{(a+1)/(a+3)} \ge 64\max\{1, Y_{\max}\}^4 D''^2$ (which is $N \ge N_B(\delta)$ along the schedule, since $c_1 \le \max\{1, Y_{\max}\}/\lambda_N$), with probability at least $1-\delta$,*

$$\|m_{\rho_N^*} - u^\dagger\|_{L^2(\nu_0)} \le \Big[24\max\{1, Y_{\max}\}Y_{\max}D' + 4C_2 + \|g_0\|\Big] N^{-\frac{a}{2(a+3)}};$$

*the rate is $N^{-1/8}$ for $a = 1$. ((**??**) with (**??**): along the schedule $c_1 \le \max\{1, Y_{\max}\}/\lambda$ and $E_* \le 2Y_{\max}/\lambda$, so $\sup|m_N - m^*| \le 24\max\{1, Y_{\max}\}Y_{\max}D'\lambda^{-3}N^{-1/2}$; balance $\lambda^{-3}N^{-1/2} = \lambda^a$.)*

Proof. The balancing: ‘λ_N^−3 N^−1/2 = λ_N^a = N^−a/(2(a+3))‘.

The high-temperature schedule: ‘β_N ≥ 1‘, ‘ϰ_N ≤ 1‘, ‘1/β_N = λ_N^a‘, ‘ϰ_N ≤ λ_N^a‘.

The bias term: ‘cor:rate-high-temperature‘ with ‘β₀ = ϰ₀ = 1‘.

The constants of ‘thm:m3-double-prime‘ along the schedule: ‘c₁ ≤ max1, Ymax/λ‘ and ‘E_* ≤ 2Ymax/λ‘, and the threshold ‘N ≥ N_B(δ)‘.

The sample term: ‘cor:m3-convergence-prime‘ (b) in the ‘$\sqrt{}$N‘ form.

The triangle inequality ‘eq:joint-triangle‘. □

Theorem 492. *Assume the setting of Theorem 328, (SC$_a$) with $0 < a \le 1$, the complexity bounds of Lemmas 314 and 315, and (S$_\infty$) for the family $\lambda \le 1$, $\beta = \beta_0$; let C_1 be Definition 145 with $\kappa_0 = 1/\beta_0$ and $M := \max\{1, Y_{\max}/\sqrt{\beta_0}\}$. With $\lambda_N := N^{-1/(2(a+3))}$, for $N^{(a+1)/(a+3)} \ge 64M^4D''^2$, with probability at least $1-\delta$,*

$$\|m_{\rho_N^*} - u^\dagger\|_{L^2(\nu_0)} \le \Big[24MY_{\max}D' + \frac{2C_1}{\beta_0} + \|g_0\|\Big] N^{-\frac{a}{2(a+3)}}.$$

(As Corollary 491, with $c_1 \le M/\lambda$ and the bias of Corollary 473. The formalized constants c_1, E_* of Theorem 381 use $M_* = Y_{\max}/\lambda$, whence the rate $N^{-a/(2(a+3))}$ in place of the paper's $N^{-a/(2(a+2))}$.)

Proof. The balancing: ‘λ_N^−3 N^−1/2 = λ_N^a = N^−a/(2(a+3))‘ and ‘λ_N ≤ λ_N^a‘.

The bias term: ‘cor:rate-sinf‘ with ‘κ_0 = 1/β_0‘.

The constants of ‘thm:m3-double-prime‘ along the schedule: ‘c_1 ≤ max1, Ymax/ $\sqrt{}$ β_0/λ‘, ‘E_* ≤ 2Ymax/λ‘, and the threshold ‘N ≥ N_B(δ)‘.

The sample term: ‘cor:m3-convergence-prime‘ (b) in the ‘ $\sqrt{}$ N‘ form.

The triangle inequality ‘eq:joint-triangle‘. □

Theorem 493. *Assume the setting of Theorem 328, (SC_a) with $0 < a \leq 1$, the complexity bounds of Lemmas 314 and 315, and (S_∞) for the family $\lambda \leq 1$, $\beta = \beta_0$ with constant B_∞; let C_1 be Definition 145 with $\kappa_0 = 1/\beta_0$, and $D' = D'(\delta)$, $D'' = D''(\delta)$ as in (**??**). Put $\lambda_1 := \min\{1, \sqrt{\beta_0}/B_\infty\}$ and $\lambda_N := N^{-1/(2(a+2))}$. Then for $N \geq N_{(i)} := \max\{\lambda_1^{-2(a+2)}, (64D''^2)^{(a+2)/a}\}$, with probability at least $1 - \delta$,*

$$\|m_{\rho_N^*} - u^\dagger\|_{L^2(\nu_0)} \leq \Big[12(B_\infty + Y_{\max})D' + \frac{2C_1}{\beta_0} + \|g_0\|\Big] N^{-\frac{a}{2(a+2)}};$$

the rate is $N^{-1/6}$ for $a = 1$ and $N^{-1/10}$ for $a = 1/2$. (For $\lambda \leq \lambda_1$, $c_1(B_\infty) = 1/\lambda$ and $E_* = B_\infty + Y_{\max}$ in Theorem 380, so $\sup |m_N - m^*| \leq 12(B_\infty + Y_{\max})D'\lambda^{-2}N^{-1/2}$*; the bias is Corollary 473; balance $\lambda^{-2}N^{-1/2} = \lambda^a$.)*

Proof. The balancing: ‘λ_N^−2 N^−1/2 = λ_N^a = N^−a/(2(a+2))‘ and ‘λ_N ≤ λ_N^a‘.

The bias term: ‘cor:rate-sinf‘ with ‘κ_0 = 1/β_0‘, and the sup norm ‘|m*| ≤ B∞‘.

The threshold ‘N ≥ N_(i)‘: ‘λ_N ≤ λ_1‘ gives ‘c_1(B∞) = 1/λ_N‘, and ‘N^a/(a+2) ≥ 64 D”^{2}‘ gives ‘N ≥ N_B(δ) = 64 c_1^{4} D”^{2}‘.

The sample term: ‘thm:m3-double-prime-rate-general‘ with ‘M_1 = B∞‘.

The triangle inequality ‘eq:joint-triangle‘. □

Theorem 494. *Assume the setting of Theorem 328, (SC_a) with $0 < a \leq 1$ and the complexity bounds of Lemmas 314 and 315; let $D' = D'(\delta)$, $D'' = D''(\delta)$ be as in (**??**). Put $c_u := \frac{1}{2}e^{\|u^\dagger\|^2/4}\|u^\dagger\|$, $\lambda_2 :=$* $\min\{1, c_u^{-2}\}$*, $\lambda_N := N^{-1/(2a+5)}$ and $\beta_N := \lambda_N^{-a}$. Then for $N \geq N_{(ii)} := \max\{\lambda_2^{-(2a+5)}, (64D''^2)^{(2a+5)/(2a+1)}\}$, with probability at least $1 - \delta$,*

$$\|m_{\rho_N^*} - u^\dagger\|_{L^2(\nu_0)} \leq \Big[12(c_u + Y_{\max})D' + 4C_2 + \|g_0\|\Big] N^{-\frac{a}{2a+5}},$$

where C_2 is Definition 147 with $\beta_0 = \kappa_0 = 1$; the rate is $N^{-1/7}$ for $a = 1$. ((SC_a) contains (R), so Lemma 467 with $\zeta_0 = \frac{\kappa}{2}\|u^\dagger\|^2 \leq \frac{1}{2}\|u^\dagger\|^2$ gives $M_* \leq c_u\lambda^{-1/2}$; for $\lambda \leq \lambda_2$ and $\beta_N \geq 1$, $c_1(M_*) = 1/\lambda$ and $E_* \leq (c_u + Y_{\max})\lambda^{-1/2}$, so $\sup |m_N - m^*| \leq 12(c_u + Y_{\max})D'\lambda^{-5/2}N^{-1/2}$*; the bias is Corollary 476 with $\beta_0 = \kappa_0 = 1$; balance $\lambda^{-5/2}N^{-1/2} = \lambda^a$.)*

Proof. The balancing: ‘λ_N^−5/2 N^−1/2 = λ_N^a = N^−a/(2a+5)‘.

The high-temperature schedule: ‘β_N ≥ 1‘, ‘κ_N ≤ 1‘, ‘1/β_N = λ_N^a‘, ‘κ_N ≤ λ_N^a‘.

The bias term: ‘cor:rate-high-temperature‘ with ‘β_0 = κ_0 = 1‘.

‘lem:sqrt-lambda-sup‘: ‘|m*| ≤ e^ζ$_0$/2‖u†‖/(2 $\sqrt{}$ λ) ≤ c_u/ $\sqrt{}$ λ‘ since ‘ζ$_0$ ≤ ‖u†‖^2/2‘.

The threshold ‘N ≥ N_(ii)‘: ‘λ_N ≤ λ_2‘ gives ‘c_1(c_u/ $\sqrt{}$ λ_N) = 1/λ_N‘, and ‘N^(2a+1)/(2a+5) ≥ 64 D”^{2}‘ gives ‘N ≥ N_B(δ) = 64 c_1^{4} D”^{2}‘.

The sample term: ‘thm:m3-double-prime-rate-general‘ with ‘M_1 = c_u/ $\sqrt{}$ λ_N‘.

The triangle inequality ‘eq:joint-triangle‘. □

Chapter 8

Dynamics: mean-field Langevin, log-Sobolev inequalities

Section 4 of the paper, restricted to what does not depend on stochastic calculus: the entropy sandwich (`lem:entropy-sandwich`), the exponential convergence of the mean-field Langevin flow from the energy identity and a uniform log-Sobolev inequality of the proximal Gibbs measures (`thm:mfld-convergence`), the decomposition, shear transport and log-Sobolev inequality of the proximal Gibbs measure (`lem:lsi`, `cor:uniform-lsi`). Of the standard facts about log-Sobolev inequalities, tensorization, Holley–Stroock, the transport by proper C^1 Lipschitz maps and the KL form (L1) are proved in the chapter "Material for Mathlib"; the Gaussian log-Sobolev inequality and the well-posedness of the dynamics are stated and left as `sorry`. The transport by a C^1 map is what the shear of `lem:lsi` needs, so `lem:lsi`(iv) assumes that the analysis maps S^*r are C^1; this is proved for a smooth feature with bounded derivatives by differentiation under the integral sign.

8.1 Entropy sandwich

Theorem 495. *For every ρ with $\int |a| \, \mathrm{d}\rho < \infty$ (in particular for $\rho \in \mathcal{D}$), $\hat{\mu}_\rho \in \mathcal{D}$.*

Proof. ☐

Theorem 496. *For ρ with $\int |a| \, \mathrm{d}\rho < \infty$ and every $\rho' \in \mathcal{D}$, $\mathrm{KL}(\rho' \| \hat{\mu}_\rho) < \infty$ and*

$$\mathcal{F}(\rho') = G_\rho(\rho') + \beta \ \mathrm{KL}(\rho' \| \hat{\mu}_\rho) - \beta \log Z_\rho, \qquad G_\rho(\rho') := L(\rho') - \int a \, s_\rho(z) \, \mathrm{d}\rho'.$$

(The chain rule $\mathrm{KL}(\rho' \| \mu_U) = \mathrm{KL}(\rho' \| \hat{\mu}_\rho) - \frac{1}{\beta} \int a \, s_\rho \, \mathrm{d}\rho' + \log \frac{Z_U}{Z_\rho}$, legitimate since $\int |a s_\rho| \, \mathrm{d}\rho' \le c_\rho \int |a| \, \mathrm{d}\rho' < \infty$.)

Proof. ☐

Theorem 497. *Assume (A3), (A4), (A5), $f \in L^2(P_X)$, and let ρ^* be the minimizer of $\mathcal{F}$ on $\mathcal{D}$. For $\rho \in \mathcal{D}$, $\mathrm{KL}(\rho \| \hat{\mu}_\rho) < \infty$ and*

$$\beta \ \mathrm{KL}(\rho \| \rho^*) \ \le \ \mathcal{F}(\rho) - \mathcal{F}(\rho^*) \ \le \ \beta \ \mathrm{KL}(\rho \| \hat{\mu}_\rho).$$

Proof. The left inequality is the strong convexity identity (Lemma 212). For the right one, write the decomposition of Lemma 496 for $\rho' = \rho$ and $\rho' = \rho^*$; by the expansion of Lemma 204 around ρ, $G_\rho(\rho^*) - G_\rho(\rho) = \frac{1}{2}\|F_{\rho^*} - F_\rho\|^2 \geq 0$, so $\mathcal{F}(\rho) - \mathcal{F}(\rho^*) = \beta\,\mathrm{KL}(\rho\|\hat{\mu}_\rho) - \beta\,\mathrm{KL}(\rho^*\|\hat{\mu}_\rho) - \frac{1}{2}\|F_{\rho^*} - F_\rho\|^2 \leq \beta\,\mathrm{KL}(\rho\|\hat{\mu}_\rho)$. □

8.2 Log-Sobolev inequality of the proximal Gibbs measures

Theorem 498. *If μ satisfies* $\mathrm{LSI}(\alpha)$ *and $0 < \alpha' \leq \alpha$, then μ satisfies* $\mathrm{LSI}(\alpha')$.

Proof. □

Theorem 499. (F1, Gross.) *For $m \in \mathbb{R}$ and $v > 0$, $\mathcal{N}(m, v)$ satisfies* $\mathrm{LSI}(1/v)$.

Proof. □

Theorem 500. (F1, Gross.) *Under (A8), $\nu_0 = \mathcal{N}(0, (\beta/\lambda_z)I)$ satisfies* $\mathrm{LSI}(\lambda_z/\beta)$.

Proof. □

Theorem 501. (F2, tensorization.) *If μ_1 satisfies* $\mathrm{LSI}(\alpha_1)$ *and μ_2 satisfies* $\mathrm{LSI}(\alpha_2)$*, then $\mu_1 \otimes \mu_2$ satisfies* $\mathrm{LSI}(\min\{\alpha_1, \alpha_2\})$.

Proof. □

Theorem 502. (F3, Holley–Stroock.) *If μ satisfies* $\mathrm{LSI}(\alpha)$*, ψ is bounded measurable and $\nu := Z^{-1}e^{-\psi}\mu$, then ν satisfies* $\mathrm{LSI}(\alpha e^{-\operatorname{osc}\psi})$*, $\operatorname{osc}\psi := \sup\psi - \inf\psi$.*

Proof. □

Theorem 503. (F4, transport.) *If μ satisfies* $\mathrm{LSI}(\alpha)$ *and T is a proper C^1 map which is Lipschitz with constant L_T, then $T_\#\mu$ satisfies* $\mathrm{LSI}(\alpha/L_T^2)$.

Proof. □

Theorem 504. *If $|s(z) - s(z')| \leq \ell\, d(z, z')$ for all z, z', then T_s and T_s^{-1} are Lipschitz with constant $1 + \ell/\lambda$ for the metric $d((a, z), (a', z')) = \max\{|a - a'|, d(z, z')\}$ on Θ.*

Proof. □

Theorem 505. *If $|s(z) - s(z')| \leq \ell\,|z - z'|$ for all z, z', then T_s^{-1} is Lipschitz with constant $1 + \ell/\lambda$ for the Euclidean metric of Θ (Lemma **??**(ii): $\|DT_s^{-1}\|_{\mathrm{op}} \leq \kappa_0(\ell/\lambda) \leq 1 + \ell/\lambda$).*

Proof. Step 1: it suffices to compare the squared norms.

Step 2: the components of ‘T⁻¹x – T⁻¹y‘ are ‘(a – a’) – (s z – s z’)/λ‘ and ‘z – z’‘.

Step 3: with ‘u = |a – a’|‘, ‘w = ‖z – z’‖‘, the first component is bounded by ‘u + k w‘, and ‘(u + k w)² + w² ≤ (1 + k)² (u² + w²)‘. □

Theorem 506. *For $\rho \in \mathcal{D}$ (indeed for any ρ with $\int |a|\,\mathrm{d}\rho < \infty$), $\hat{\mu}_\rho(\mathrm{d}a\,\mathrm{d}z) = \hat{\nu}_\rho(\mathrm{d}z)\,\mathcal{N}(-s_\rho(z)/\lambda, \beta/\lambda)(\mathrm{d}a)$, where $\hat{\nu}_\rho$ is the hidden marginal of $\hat{\mu}_\rho$.*

Proof. □

Theorem 507. *The hidden marginal of $\hat{\mu}_\rho$ is $\hat{\nu}_\rho(\mathrm{d}z) = \hat{Z}_\rho^{-1}\exp(\frac{s_\rho(z)^2}{2\lambda\beta})\nu_0(\mathrm{d}z)$, with $\hat{Z}_\rho = \int e^{s_\rho^2/(2\lambda\beta)}\,\mathrm{d}\nu_0 \in [1, e^{c_\rho^2/(2\lambda\beta)}]$.*

Proof. ☐

Theorem 508. *For $T_\rho(a,z) := (a + s_\rho(z)/\lambda,\ z)$, $(T_\rho)_\# \hat\mu_\rho = \mathcal{N}(0,\beta/\lambda) \otimes \hat\nu_\rho$.*

Proof. By Lemma 506, under $\hat\mu_\rho$ one has $z \sim \hat\nu_\rho$ and $a \mid z \sim \mathcal{N}(-s_\rho(z)/\lambda, \beta/\lambda)$, so $a + s_\rho(z)/\lambda \mid z \sim \mathcal{N}(0,\beta/\lambda)$ does not depend on z. ☐

Theorem 509. *If $z \mapsto \varphi_z(x)$ is C^1 for every x and Z is finite dimensional, then $x \mapsto D_z\varphi_z(x)$ is measurable for every z (a limit of difference quotients along a basis).*

Proof. Step 1: the difference quotients along the basis are measurable in ‘x‘.
Step 2: a linear form is the sum of its values on the basis times the coordinates.
Step 3: the difference quotients converge to the derivative. ☐

Theorem 510. *If the feature map is smooth with bounded derivatives (Definition 127), Z is finite dimensional and $r \in L^1(P_X)$, then $S^*r\colon z \mapsto \int \varphi_z(x) r(x) P_X(\,\mathrm{d}x)$ is C^1 with $D(S^*r)(z) = \int r(x) D_z\varphi_z(x) P_X(\,\mathrm{d}x)$.*

Proof. Differentiation under the integral sign: $|r(x) D_z\varphi_z(x)| \le L|r(x)|$ with $L = \sup\|\nabla_z\varphi\|$ is an integrable bound, $x \mapsto D_z\varphi_z(x)$ is measurable (Lemma 509), and the derivative is continuous by dominated convergence. ☐

Theorem 511. *Under (A8), $\hat\nu_\rho$ satisfies $\mathrm{LSI}(\hat\alpha_\rho)$ with $\hat\alpha_\rho \ge \frac{\lambda_z}{\beta}\exp(-\frac{\mathrm{osc}(s_\rho^2)}{2\lambda\beta}) \ge \frac{\lambda_z}{\beta}\exp(-\frac{c_\rho^2}{2\lambda\beta})$.*

Proof. $\nu_0 = \mathcal{N}(0,(\beta/\lambda_z)I)$ satisfies $\mathrm{LSI}(\lambda_z/\beta)$ by (F1), and $\hat\nu_\rho = Z^{-1}e^{-\psi}\nu_0$ with $\psi = -s_\rho^2/(2\lambda\beta)$, $-c_\rho^2/(2\lambda\beta) \le \psi \le 0$; apply (F3). ☐

Theorem 512. *Under (A1), (A3), (A5), (A8), and assuming that S^*r is C^1 for every $r \in L^2(P_X)$ (Lemma 510), $\hat\mu_\rho$ satisfies $\mathrm{LSI}(\alpha_\rho)$ with*

$$\alpha_\rho \;\ge\; \frac{\min\{\lambda,\ \lambda_z e^{-c_\rho^2/(2\lambda\beta)}\}}{\beta\,(1+\ell_\rho/\lambda)^2} \;\ge\; \frac{\min\{\lambda,\ \lambda_z e^{-c_\rho^2/(2\lambda\beta)}\}}{\beta\,(1+\sqrt{R_X^2+1}\,c_\rho/\lambda)^2},$$

where $\ell_\rho = \sup|\nabla s_\rho| \le \sqrt{R_X^2+1}\,c_\rho$. The right-hand side does not depend on the dimension m.

Proof. $\mathcal{N}(0,\beta/\lambda)$ satisfies $\mathrm{LSI}(\lambda/\beta)$ (F1), so by (F2) and Lemma 511 the product $\pi_\rho = \mathcal{N}(0,\beta/\lambda)\otimes\hat\nu_\rho$ satisfies $\mathrm{LSI}(\min\{\lambda/\beta, \hat\alpha_\rho\})$. By Lemma 508, $\hat\mu_\rho = (T_\rho^{-1})_\#\pi_\rho$, and T_ρ^{-1} is Lipschitz with constant $1+\ell_\rho/\lambda$ (Lemma 505, with $\ell_\rho \le \sqrt{R_X^2+1}\,c_\rho$ by Lemma 191); it is C^1 since s_ρ is, and proper since it is a homeomorphism (its inverse is T_ρ); conclude by (F4). ☐

Theorem 513. *For $\rho \in \mathcal{D}$, $c_\rho^2 = 2L(\rho) \le 2(\mathcal{F}(\rho) + \beta\log Z_U)$ (here $c_\rho^2 \le 2\mathcal{F}(\rho)$, the constant of the free energy being dropped).*

Proof. ☐

Theorem 514. *The map $c \mapsto \alpha(c)$ of Definition 125 is nonincreasing on $[0,\infty)$.*

Proof. ☐

Theorem 515. *For $E \in \mathbb{R}$, on the sublevel set $\mathcal{D}_E := \{\rho \in \mathcal{D} : \mathcal{F}(\rho) \le E\}$ and with $c_E := \sqrt{2(E+\beta\log Z_U)}$,*

$$\inf_{\rho\in\mathcal{D}_E} \alpha_\rho \;\ge\; \alpha_*(E) = \frac{\min\{\lambda,\ \lambda_z\, e^{-c_E^2/(2\lambda\beta)}\}}{\beta\,(1+\sqrt{R_X^2+1}\,c_E/\lambda)^2} > 0,$$

i.e. $\hat\mu_\rho$ satisfies $\mathrm{LSI}(\alpha_*(E))$ for every $\rho \in \mathcal{D}_E$.

Proof. $c_\rho \le c_E$ on $\mathcal{D}_E$ by Corollary 513, and the constant of Lemma 512 is nonincreasing in c_ρ (Lemma 514); conclude by Lemma 498. $\square$

Theorem 516. *For $\rho_0 = \mu_U = \mathcal{N}(0, \beta/\lambda) \otimes \nu_0$, $F_{\mu_U} = 0$, $L(\mu_U) = \|f\|^2/2$ and $\mathcal{F}(\mu_U) = \|f\|^2/2 - \beta \log Z_U$ (here $\|f\|^2/2$).*

Proof. Step 1: 'F_μ_U(x) = (∫ a d𝒩(0, β/λ)) (∫ φ_z(x) dν₀) = 0'.
Step 2: 'L(μ_U) = ½ ∫ f² dP = ½ ‖f‖²'.
Step 3: 'KL(μ_U‖μ_U) = 0'. $\square$

Theorem 517. *For $\rho_0 = \mu_U$, $c_{\mathcal{F}(\mu_U)} = \|f\|$ and*

$$\alpha_* := \alpha_*(\mathcal{F}(\mu_U)) = \frac{\min\{\lambda,\ \lambda_z\, e^{-\|f\|^2/(2\lambda\beta)}\}}{\beta\,(1 + \sqrt{R_X^2 + 1}\,\|f\|/\lambda)^2}.$$

Proof. $\square$

Theorem 518. *Since $\rho^* \in \mathcal{D}_{\mathcal{F}(\mu_U)}$, $\rho^* = \hat{\mu}_{\rho^*}$ satisfies* $\mathrm{LSI}(\alpha_*)$.

Proof. $\square$

8.3 Exponential convergence of the mean-field Langevin dynamics

Theorem 519. (L1.) *If μ satisfies $\mathrm{LSI}(\alpha)$ on a finite-dimensional space, $\rho \ll \mu$ is a probability measure with $\mathrm{KL}(\rho\|\mu) < \infty$, $\log \frac{\mathrm{d}\rho}{\mathrm{d}\mu}$ is C^1 and $\nabla \log \frac{\mathrm{d}\rho}{\mathrm{d}\mu} \in L^2(\rho)$, then $\mathrm{KL}(\rho\|\mu) \le \frac{1}{2\alpha} I(\rho|\mu)$ (substitute $g = \sqrt{\mathrm{d}\rho/\mathrm{d}\mu}$, cut off by compactly supported functions, in the LSI).*

Proof. $\square$

Theorem 520. *If a probability measure μ on Θ (Z finite dimensional) satisfies $\mathrm{LSI}(\alpha)$, then it satisfies the KL form of $\mathrm{LSI}(\alpha)$.*

Proof. $\square$

Theorem 521. *Along an MFLD flow, $t \mapsto \mathcal{F}(\rho_t)$ is nonincreasing on $[0, \infty)$.*

Proof. $\square$

Theorem 522. *Assume (A3), (A5), $f \in L^2(P_X)$, let $(\rho_t)_{t\ge 0}$ be an MFLD flow (Hypothesis W) and assume that $\hat{\mu}_{\rho_t}$ satisfies the KL form of $\mathrm{LSI}(\alpha_*)$ for every $t \ge 0$. Then for all $t \ge 0$*

$$\mathcal{F}(\rho_t) - \mathcal{F}(\rho^*) \le e^{-2\alpha_*\beta t}(\mathcal{F}(\rho_0) - \mathcal{F}(\rho^*)).$$

Proof. Put $\Phi(t) := \mathcal{F}(\rho_t) - \mathcal{F}(\rho^*) \ge 0$. By (W2), Φ is nonincreasing. For a.e. $u \ge 0$, the right inequality of Lemma 497 and (L1) give $\Phi(u) \le \beta\, \mathrm{KL}(\rho_u\|\hat{\mu}_{\rho_u}) \le \frac{\beta}{2\alpha_*} I(\rho_u|\hat{\mu}_{\rho_u})$. Hence, for $0 \le s \le t$, $\Phi(t) = \Phi(s) - \beta^2 \int_s^t I(\rho_u|\hat{\mu}_{\rho_u})\,\mathrm{d}u \le \Phi(s) - 2\alpha_*\beta \int_s^t \Phi(u)\,\mathrm{d}u \le \Phi(s) - 2\alpha_*\beta(t-s)\Phi(t)$, and Lemma 909 (Grönwall) concludes. $\square$

Theorem 523. *Under the hypotheses of Theorem 522, for all $t \ge 0$, $\mathrm{KL}(\rho_t\|\rho^*) \le \frac{1}{\beta} e^{-2\alpha_*\beta t}(\mathcal{F}(\rho_0) - \mathcal{F}(\rho^*))$. (The left inequality of Lemma 497.)*

Proof. □

Theorem 524. *Under the hypotheses of Theorem 522, for all $t \geq 0$, $\|F_{\rho_t} - F_{\rho^*}\|^2_{L^2(P_X)} \leq 2e^{-2\alpha_*\beta t}(\mathcal{F}(\rho_0) - \mathcal{F}(\rho^*))$. (The strong convexity identity of Lemma 212.)*

Proof. □

Theorem 525. *Assume (A1), (A3), (A5), (A8), $f \in L^2(P_X)$ and (A9): $\rho_0 \in \mathcal{D}$, $\int |z|^2 \, \mathrm{d}\rho_0 < \infty$ and $\int e^{c_0|a|} \, \mathrm{d}\rho_0 < \infty$ for some $c_0 > 0$. Then the MFLD has a solution on $[0, \infty)$ (unique in law), and its flow of laws $(\rho_t)_{t \geq 0}$ satisfies (W1) and (W2), i.e. it is an MFLD flow with ρ_0 as initial law.*

Proof. □

Theorem 526. *Assume (A1), (A3), (A5), (A8), $f \in L^2(P_X)$ and (A9). Let $(\rho_t)_{t \geq 0}$ be the flow of laws of the solution of Theorem 525 and $\alpha_* := \alpha_*(\mathcal{F}(\rho_0))$. Then for all $t \geq 0$*

$$\mathcal{F}(\rho_t) - \mathcal{F}(\rho^*) \leq e^{-2\alpha_*\beta t}(\mathcal{F}(\rho_0) - \mathcal{F}(\rho^*)), \quad \mathrm{KL}(\rho_t \| \rho^*) \leq \frac{1}{\beta} e^{-2\alpha_*\beta t}(\mathcal{F}(\rho_0) - \mathcal{F}(\rho^*)),$$

$$\|F_{\rho_t} - F_{\rho^*}\|^2_{L^2(P_X)} \leq 2e^{-2\alpha_*\beta t}(\mathcal{F}(\rho_0) - \mathcal{F}(\rho^*)).$$

Proof. Theorem 525 gives the MFLD flow; since $\mathcal{F}(\rho_t) \leq \mathcal{F}(\rho_0)$ (Lemma 521), Corollary 515 gives $\mathrm{LSI}(\alpha_*)$ for every $\hat{\mu}_{\rho_t}$ (the analysis maps are C^1 by Lemma 510, the feature being smooth with bounded derivatives), hence its KL form (Lemma 520), and Theorems 522, 523, 524 apply. □

8.4 Static chaos of the finite-particle Gibbs measure

The M-particle free energy decomposes exactly relative to the product of the mean-field minimizer, so that the relative entropy of the M-particle Gibbs measure with respect to $\rho^{*\otimes M}$ is bounded uniformly in M; this closes the order "$t \to \infty$ at fixed M first" of the order-of-limits theorem under the ergodicity of the finite-particle dynamics (manuscript note `notes-14`).

Theorem 527. *Let μ be a nonzero σ-finite measure on Θ, $\beta > 0$ and W measurable with $Z_W = \int e^{-W/\beta} \, \mathrm{d}\mu < \infty$. Then $\hat{\mu}_W^{\otimes M} = \widehat{(\mu^{\otimes M})}_{W_M}$ with $W_M(\theta) = \sum_\ell W(\theta_\ell)$, i.e. $\frac{\mathrm{d}\hat{\mu}_W^{\otimes M}}{\mathrm{d}\mu^{\otimes M}}(\theta) = \prod_\ell \frac{e^{-W(\theta_\ell)/\beta}}{Z_W} = \frac{e^{-W_M(\theta)/\beta}}{Z_W^M}$, and $Z_{W_M} = Z_W^M$.*

Proof. Step 1: 'e^−W_M/β = $\prod_i$ e^−W(θ$_i$)/β' and 'Z_W_M = Z_W^|ι|'.
Step 2: both sides are 'μ^⊗ι' with the product density ('lem:pi-withDensity'). □

Theorem 528. *For every $\theta \in \Theta^M$, $0 \leq L(\rho_\theta) \leq \frac{1}{2}\left(\frac{1}{M}\sum_\ell |a_\ell| + \|f\|\right)^2 \leq \frac{1}{M}\sum_\ell a_\ell^2 + \|f\|^2$.*

Proof. □

Theorem 529. *Let $\mu \in \mathcal{P}(\Theta^M)$ with $\mathrm{KL}(\mu \| \mu_U^{\otimes M}) < \infty$. Then $\int a_\ell^2 \, \mu(\mathrm{d}\theta) < \infty$ for every ℓ, hence $\int |a_\ell| \, \mathrm{d}\mu < \infty$ and $L(\rho_\theta)$, $\sum_\ell a_\ell s(z_\ell)$ (s bounded measurable) are μ-integrable. (Donsker–Varadhan with $g(\theta) = \frac{\lambda}{4\beta}a_\ell^2$, $\int e^g \, \mathrm{d}\mu_U^{\otimes M} = \int e^{\lambda a^2/(4\beta)} \, \mathrm{d}\mu_U = \sqrt{2}$.)*

Proof. □

Theorem 530. *Since $0 \leq L(\rho_\theta)$, $Z_M = \int e^{-ML(\rho_\theta)/\beta} \, \mathrm{d}\mu_U^{\otimes M} \in (0, 1]$, so π_M is a probability measure, and $\mathrm{KL}(\pi_M \| \mu_U^{\otimes M}) = -\frac{M}{\beta}\int L(\rho_\theta) \, \mathrm{d}\pi_M - \log Z_M < \infty$ (the potential $ML(\rho_\theta)$ is π_M-integrable because $ue^{-u/\beta} \leq \beta$).*

Proof. □

Theorem 531. *For $\mu \in \mathcal{P}(\Theta^M)$ with $\mathrm{KL}(\mu\|\mu_U^{\otimes M}) < \infty$, $\mathrm{KL}(\mu\|\pi_M) < \infty$ and $\mathcal{F}^M(\mu) = \beta\ \mathrm{KL}(\mu\|\pi_M) - \beta \log Z_M$, $Z_M := \int e^{-ML(\rho_\theta)/\beta}\,\mathrm{d}\mu_U^{\otimes M}$ (Lemma 187 with $W = ML(\rho_\theta)$ and reference measure $\mu_U^{\otimes M}$; $W \in L^1(\mu)$ by Lemma 529).*

Proof. □

Theorem 532. *π_M is the minimizer of $\mathcal{F}^M$ on $\{\mu \in \mathcal{P}(\Theta^M) : \mathrm{KL}(\mu\|\mu_U^{\otimes M}) < \infty\}$: $\mathcal{F}^M(\pi_M) = -\beta \log Z_M \le \mathcal{F}^M(\mu)$.*

Proof. □

Theorem 533. *With $s^* = S^* r_{\rho^*}$ and $W^*(\theta) = a\, s^*(z)$, $\rho^{*\otimes M} = \widehat{(\mu_U^{\otimes M})}_{W_M^*}$, $W_M^*(\theta) = \sum_\ell a_\ell s^*(z_\ell) = M \int a s^*\,\mathrm{d}\rho_\theta$, and $\log \frac{\mathrm{d}\rho^{*\otimes M}}{\mathrm{d}\mu_U^{\otimes M}}(\theta) = -\frac{1}{\beta}\sum_\ell a_\ell s^*(z_\ell) - M \log Z_*$.*

Proof. □

Theorem 534. *Let ρ^* be the minimizer of $\mathcal{F}$ on $\mathcal{D}$. For every $\mu \in \mathcal{P}(\Theta^M)$ with $\mathrm{KL}(\mu\|\mu_U^{\otimes M}) < \infty$, $\mathrm{KL}(\mu\|\rho^{*\otimes M}) < \infty$ and*

$$\mathcal{F}^M(\mu) - M\mathcal{F}(\rho^*) = \beta\ \mathrm{KL}(\mu\|\rho^{*\otimes M}) + \frac{M}{2}\int_{\Theta^M} \|F_{\rho_\theta} - F_{\rho^*}\|^2_{L^2(P_X)}\,\mu(\,\mathrm{d}\theta).$$

(Chain rule for $\rho^{*\otimes M} = \widehat{(\mu_U^{\otimes M})}_{W_M^*}$, the identity $\mathcal{F}(\rho^*) = L(\rho^*) - \int a s^*\,\mathrm{d}\rho^* - \beta \log Z_*$ and the expansion $L(\rho_\theta) - L(\rho^*) = \int a s^*\,\mathrm{d}(\rho_\theta - \rho^*) + \frac{1}{2}\|F_{\rho_\theta} - F_{\rho^*}\|^2$.)

Proof. Step 0: notation ‘s* = S* r_ρ*‘, ‘W(θ) = a s*(z)‘, ‘W_M(θ) = ∑_ℓ W(θ_ℓ)‘.
Step 1: ‘Z_W_M = Z_*^M‘.
Step 2: the chain rule ‘KL(μ‖ρ*^⊗M) = KL(μ‖μ_U^⊗M) + (1/β) ∫ W_M dμ + M log Z_*‘.
Step 3: ‘ℱ(ρ*) = L(ρ*) − ∫ W dρ* − β log Z_*‘.
Step 4: the pointwise expansion ‘M L(ρ_θ) − W_M(θ) − M (L(ρ*) − ∫ W dρ*) = (M/2) ‖F_ρ_θ − F_ρ*‖²‘.
Step 5: integrate the pointwise identity against ‘μ‘. □

Theorem 535. *Let $\rho \in \mathcal{P}(\Theta)$ with $\int a^2\,\mathrm{d}\rho < \infty$ and $M \ge 1$. Then*

$$\int_{\Theta^M} \|F_{\rho_\theta} - F_\rho\|^2_{L^2(P_X)}\,\rho^{\otimes M}(\,\mathrm{d}\theta) = \frac{1}{M}\int_{\mathcal{X}} \mathrm{Var}_\rho(a\varphi_z(x))\,P_X(\,\mathrm{d}x) \le \frac{1}{M}\int a^2\,\mathrm{d}\rho.$$

(For each x, $F_{\rho_\theta}(x) = \frac{1}{M}\sum_\ell a_\ell \varphi_{z_\ell}(x)$ is a mean of i.i.d. variables of mean $F_\rho(x)$; Fubini is justified by $\int a^2\,\mathrm{d}\rho < \infty$.)

Proof. Step A: ‘‖F_ρ_θ − F_ρ‖² = ∫ (F_ρ_θ(x) − F_ρ(x))² P(dx)‘ for every ‘θ‘.
Step B: the integrand is dominated by ‘2 (1/M) ∑_ℓ a_ℓ² + 2 (∫ |a| dρ)²‘, so Fubini applies.
Step C: for each ‘x‘, the variance of the mean of ‘M‘ i.i.d. copies of ‘a φ_z(x)‘ is ‘(1/M) Var_ρ(a φ_z(x)) ≤ (1/M) ∫ a² dρ‘.
Step D: Fubini and integration of the pointwise bound over ‘x‘. □

Theorem 536. $\mathrm{KL}(\rho^{*\otimes M}\|\mu_U^{\otimes M}) = M\ \mathrm{KL}(\rho^*\|\mu_U) < \infty$.

Proof. □

Theorem 537. *Let ρ^* be the minimizer of $\mathcal{F}$ on $\mathcal{D}$ and $M \geq 1$. Then*

$$\beta \ \mathrm{KL}(\pi_M \| \rho^{*\otimes M}) + \frac{M}{2} \mathbb{E}_{\pi_M} \| F_{\rho_\theta} - F_{\rho^*} \|^2_{L^2(P_X)} \leq \frac{1}{2} \int a^2 \, \mathrm{d}\rho^*.$$

(Theorem 534 at $\mu = \pi_M$ and at $\mu = \rho^{*\otimes M}$, the minimality of π_M for $\mathcal{F}^M$, and Lemma 535.)

Proof. Step 1: the identity at ‘μ = π_M‘ and at ‘μ = ρ*^⊗M‘.
Step 2: minimality of ‘π_M‘ and the i.i.d. variance bound. □

Theorem 538. *Under the hypotheses of Theorem 537, $\mathrm{KL}(\pi_M \| \rho^{*\otimes M}) \leq \frac{1}{2\beta} \int a^2 \, \mathrm{d}\rho^*$ (independent of M) and $\mathbb{E}_{\pi_M} \| F_{\rho_\theta} - F_{\rho^*} \|^2 \leq \frac{1}{M} \int a^2 \, \mathrm{d}\rho^*$.*

Proof. □

Theorem 539. *With $A := \min\{\|f\|^2/\lambda^2, \|f\|^2/(4\lambda)\} + \beta/\lambda \geq \int a^2 \, \mathrm{d}\rho^*$ (Theorem 233), $\mathrm{KL}(\pi_M \| \rho^{*\otimes M}) \leq A/(2\beta)$ and $\mathbb{E}_{\pi_M} \| F_{\rho_\theta} - F_{\rho^*} \|^2 \leq A/M$.*

Proof. □

Theorem 540. *Let s be measurable with $|s| \leq C$ and $\rho_s = \hat{\mu}_{W_s}$, $W_s(\theta) = a\, s(z)$. Then $\int e^{c|a|} \, \mathrm{d}\rho_s < \infty$ for every $c \in \mathbb{R}$, since $e^{-as(z)/\beta} e^{c|a|} \leq e^{(C/\beta + c)|a|}$ and the a-marginal of μ_U is Gaussian.*

Proof. □

Theorem 541. *Under the hypotheses of Lemma 540, for $c \geq 0$, $\int e^{c|a|} \, \mathrm{d}\rho_s \leq Z_s^{-1} \int e^{(c + C/\beta)|a|} \, \mathrm{d}\mu_U \leq 2 \exp(\frac{\beta}{2\lambda}(c + C/\beta)^2)$, using $Z_s \geq 1$ and the Gaussian moment generating function.*

Proof. Step 1: ‘∫ e^c|a| dρ_s = Z⁻¹ ∫ e^−a s/β e^c|a| dμ_U ≤ Z⁻¹ ∫ e^t|a| dμ_U‘.
Step 2: ‘∫ e^t|a| dμ_U ≤ ∫ (e^ta + e^−ta) dμ_U = 2 e^t² (β/λ)/2‘ (Gaussian moment generating function). □

Theorem 542. *Let g be measurable with $|g| \leq C$, $m := \int ag(z) \, \mathrm{d}\rho^*$ and $Y(\theta) := ag(z) - m$. Then $K_g := \int Y^2 e^{|Y|} \, \mathrm{d}\rho^* < \infty$, since $|Y| \leq C|a| + |m|$, $u^2 \leq 2e^u$ and $\int e^{2C|a|} \, \mathrm{d}\rho^* < \infty$.*

Proof. □

Theorem 543. *Under the hypotheses of Lemma 542, $K_g \leq 4 \exp(2C(\|f\|/\lambda + \sqrt{2\beta/(\pi\lambda)})) \exp(\frac{\beta}{2\lambda}(2C + \|f\|/\beta)^2)$, a constant depending only on $C, \|f\|, \lambda, \beta$.*

Proof. Step 1: ‘|m| ≤ C ∫ |a| dρ* ≤ C (‖f‖/λ + √(2β/(πλ)))‘.
Step 2: the pointwise bound ‘Y² e^|Y| ≤ 2 e^2|m| e^2C|a|‘.
Step 3: integrate and use the Gaussian exponential-moment bound. □

Theorem 544. *Let $g \colon Z \to \mathbb{R}$ be measurable with $|g| \leq C$, $m := \int ag(z) \, \mathrm{d}\rho^* = \int g \, \mathrm{d}\Pi\rho^*$, $\bar{X}(\theta) := \frac{1}{M} \sum_\ell a_\ell g(z_\ell) = \int g \, \mathrm{d}\Pi\rho_\theta$ and $K_g := \int (ag(z) - m)^2 e^{|ag(z) - m|} \, \mathrm{d}\rho^* < \infty$. Then*

$$\mathbb{E}_{\pi_M} \Big| \int g \, \mathrm{d}\Pi\rho_\theta - \int g \, \mathrm{d}\Pi\rho^* \Big| \leq \frac{\mathrm{KL}(\pi_M \| \rho^{*\otimes M}) + \log 2 + K_g}{\sqrt{M}} \leq \frac{C_g}{\sqrt{M}}, \qquad C_g := \frac{A}{2\beta} + \log 2 + K_g.$$

(Donsker–Varadhan on Θ^M with $h = \sqrt{M}|\bar{X} - m|$: with $u = 1/\sqrt{M}$ and $Y_\ell = a_\ell g(z_\ell) - m$ i.i.d. centered, $\mathbb{E}_{\rho^{*\otimes M}} e^{\pm u \sum_\ell Y_\ell} = (\mathbb{E} e^{\pm uY})^M \leq (1 + K_g u^2)^M \leq e^{K_g}$, so $\log \mathbb{E}_{\rho^{*\otimes M}} e^h \leq \log 2 + K_g$.)

Proof. Step 0: notation. ‘Y(p) = a g(z) − m‘ is centered with ‘K = ∫ Y² e^|Y| dρ* < ∞‘; ‘S(θ) = ∑_ℓ Y(θ_ℓ)‘, ‘u = 1/√M‘, ‘h(θ) = u |S(θ)| = √M |X̄(θ) − m|‘.

Step 1: ‘X̄(θ) − m = (1/M) S(θ)‘.
Step 2: the exponential moments of ‘± u S‘ under ‘ρ*^⊗M‘ are at most ‘e^K‘.
Step 3: ‘e^u|S| ≤ e^uS + e^−uS‘, so ‘∫ e^u|S| dρ*^⊗M ≤ 2 e^K‘.
Step 4: Donsker–Varadhan on ‘Θ^M‘: ‘∫ u|S| dπ_M ≤ KL(π_M‖ρ*^⊗M) + log 2 + K‘.
Step 5: ‘|X̄ − m| = u · (u |S|)‘ since ‘u² = 1/M‘; conclude. □

Theorem 545. *Under the hypotheses of Proposition 544, with $A = \min\{\|f\|^2/\lambda^2, \|f\|^2/(4\lambda)\} + \beta/\lambda$, $\mathbb{E}_{\pi_M}|\bar{X} - m| \le C_g/\sqrt{M}$ with $C_g = A/(2\beta) + \log 2 + K_g$.*

Proof. □

Theorem 546. *Under the hypotheses of Proposition 544, $\mathbb{E}_{\pi_M}|\bar{X} - m| \le C_g/\sqrt{M}$ with the explicit constant $C_g = \frac{A}{2\beta} + \log 2 + 4\exp(2C(\|f\|/\lambda + \sqrt{2\beta/(\pi\lambda)}))\exp(\frac{\beta}{2\lambda}(2C + \|f\|/\beta)^2)$, which depends only on $C, \|f\|, \lambda, \beta$.*

Proof. □

Theorem 547. *For every permutation σ of $\{1, \dots, M\}$, the image of π_M under $\theta \mapsto \theta \circ \sigma$ is π_M: $\mu_U^{\otimes M}$ is permutation invariant and $L(\rho_{\theta\circ\sigma}) = L(\rho_\theta)$. In particular all one-particle marginals of π_M coincide.*

Proof. □

Theorem 548. *Let Z be standard Borel. For every ℓ, the one-particle marginal $\pi_M^{(\ell)}$ of π_M satisfies $M\ \mathrm{KL}(\pi_M^{(\ell)}\|\rho^*) = \sum_{\ell'} \mathrm{KL}(\pi_M^{(\ell')}\|\rho^*) \le \mathrm{KL}(\pi_M\|\rho^{*\otimes M})$, by exchangeability and the superadditivity $\mathrm{KL}(\mu\|\nu^{\otimes M}) \ge \sum_\ell \mathrm{KL}(\mu^{(\ell)}\|\nu)$.*

Proof. □

Theorem 549. *Under the hypotheses of Theorem 537 with Z standard Borel, $\mathrm{KL}(\pi_M^{(\ell)}\|\rho^*) \le \frac{A}{2\beta M}$ for every ℓ.*

Proof. □

Theorem 550. *Assume (E): $\|\mu_t - \pi_M\|_{\mathrm{TV}} \to 0$ and $\sup_t \int \frac{1}{M}\sum_\ell a_\ell^2\, \mathrm{d}\mu_t \le Q < \infty$. Then for every bounded measurable g, with $\bar{X}(\theta) = \int g\, \mathrm{d}\Pi\rho_\theta$, $m = \int g\, \mathrm{d}\Pi\rho^*$ and C_g as in Proposition 544,*

$$\limsup_{t\to\infty} \mathbb{E}_{\mu_t}|\bar{X} - m| \le \frac{C_g}{\sqrt{M}},$$

hence $\limsup_{M\to\infty} \limsup_{t\to\infty} \mathbb{E}\left|\int g\, \mathrm{d}\Pi\rho_t^M - \int g\, \mathrm{d}\Pi\rho^*\right| = 0$. (Truncation: $h \wedge T$ converges by total variation, and $\mathbb{E}_{\mu_t}[h - h\wedge T] \le \mathbb{E}_{\mu_t} h^2/T \le (2C^2Q + 2m^2)/T$ uniformly in t.)

Proof. Step 0: notation: ‘m‘, the statistic ‘h(θ) = |X̄(θ) − m|‘, its bound ‘C S(θ) + |m|‘ with ‘S(θ) = (1/M) ∑_ℓ |a_ℓ|‘, the bound ‘B‘ of Proposition 14.4 and the second-moment bound ‘Q_2 = 2 C² Q + 2 m²‘ of ‘h‘.

Step 1: the truncation level ‘T‘ with ‘Q_2/T ≤ ε/2‘.
Step 2: integrability of ‘h‘, ‘h²‘ and ‘h ∧ T‘ under each ‘μ_t‘ and under ‘π_M‘.
Step 3: the tail bound ‘∫ h dμ_t ≤ ∫ h ∧ T dμ_t + Q_2/T‘ (Chebyshev: ‘h − h ∧ T ≤ h²/T‘).
Step 4: total-variation convergence for the truncated statistic, and ‘∫ h ∧ T dπ_M ≤ B‘.
Step 5: eventually ‘2 T ‖μ_t − π_M‖_TV ≤ ε/2‘. □

Theorem 551. *For the tanh feature (A3) with $z = (w, b) \in \mathbb{R}^{m+1}$ (Euclidean norm), $\|\nabla_z \varphi_z(x)\| \leq \sqrt{|x|^2 + 1}$ for all $x \in \mathbb{R}^m$ and $z \in \mathbb{R}^{m+1}$. Indeed $\nabla_z \varphi_z(x) = \tanh'(w^\top x - b)\,(x, -1)$ with $0 < \tanh' \leq 1$ and $|(x, -1)| = \sqrt{|x|^2 + 1}$. In particular, under (A1) the feature is Lipschitz in the hidden parameter with constant $\sqrt{R_X^2 + 1}$.*

Proof. Step 1: the chain rule bound '‖∇_z φ_z(x)‖ ≤ |tanh'(⟨⟨w, x⟩⟩ − b)| · ‖(x, −1)‖'.

Step 2: '|tanh'| = 1 − tanh² ≤ 1' and '‖(x, −1)‖ = √(‖x‖² + 1)'. □

Theorem 552. *For the tanh feature (A3), with $z = (w, b) \in \mathbb{R}^{m+1}$ (Euclidean norm, Definition 6), and inputs $|x| \leq R_X$ (A1), the map $z \mapsto \varphi_z(x)$ is C^∞ and, for every $n \geq 0$,*

$$\sup_{|x| \leq R_X,\, z} \|\nabla_z^n \varphi_z(x)\| \leq \|\tanh^{(n)}\|_\infty \, (R_X^2 + 1)^{n/2} < \infty,$$

where $\|\tanh^{(n)}\|_\infty \leq \sup_{|s| \leq 1} |p_n(s)|$ for the polynomials $p_0 = X$, $p_{n+1} = p_n' \,(1 - X^2)$ with $\tanh^{(n)} = p_n(\tanh)$. Hence the restriction of the tanh feature to $\{|x| \leq R_X\}$ (Definition 3) is smooth with bounded derivatives in the sense of Definition 127.

Proof. Step 1: smoothness. 'z ↦ tanh(⟨⟨w, x⟩⟩ − b)' is the composition of the 'C^∞' function 'tanh' with the continuous linear map '(w, b) ↦ ⟨⟨(x, −1), (w, b)⟩⟩'.

Step 2: the 'n'-th derivative of 'tanh' is bounded by a constant 'c_n' ('tanh$^{(n)}$ = p_n(tanh)' with '|tanh| ≤ 1').

Step 3: the chain rule for the linear map '(w, b) ↦ ⟨⟨(x, −1), (w, b)⟩⟩' of norm '√(‖x‖² + 1) ≤ √(R_X² + 1)' gives '‖∇_z^n φ_z(x)‖ ≤ c_n (R_X² + 1)^n/2'. □

8.5 The third stage of the order of the limits: $t \to \infty$ at fixed M first

Theorem 553. *If $\int |a| \, \mathrm{d}\rho < \infty$ and g is bounded measurable, then $\int_Z g \, \mathrm{d}\Pi\rho = \int_\Theta a\, g(z)\, \rho(\mathrm{d}a \, \mathrm{d}z)$.*

Proof. □

Theorem 554. *Fix the sample $(x_i, y_i)_{i=1}^N$ with $|y_i| \leq Y_{\max}$, (λ, β) and $M \geq 1$; let ρ_N^* be the minimizer of $\mathcal{F}_N$ and let $(\mu_t)_{t \geq 0}$ be ergodic to the empirical M-particle Gibbs measure π_M (Assumption (E)). For every bounded measurable g with $\|g\|_\infty \leq C$ and every $\varepsilon > 0$ there is t_ε such that for $t \geq t_\varepsilon$*

$$\mathbb{E}_{\mu_t} e_{M,t}^{(g)} = \mathbb{E}_{\mu_t} \left| \int g \, \mathrm{d}\Pi\rho_\theta - \int g \, \mathrm{d}\Pi\rho_N^* \right| \leq \frac{C_g(Y_{\max}, \lambda, \beta)}{\sqrt{M}} + \varepsilon,$$

with the constant of Definition 135, which depends on the sample only through $Y_{\max}$. (Theorem 550 for the empirical feature, $\|y\|_N \leq Y_{\max}$, and $\int g \, \mathrm{d}\Pi\rho_N^* = \int a\, g(z) \, \mathrm{d}\rho_N^*$.)

Proof. Step 1: the dictionary: 'ρ_N^*' minimizes the free energy of the empirical feature with target 'y ∈ L²(uniform)', and '‖y‖_N ≤ Ymax'.

Step 2: 'thm:static-chaos-order' for the empirical feature.

Step 3: the constant is bounded uniformly in the sample. □

Theorem 555. *Assume the hypotheses of Theorem 427 and let, for every sample and every $M \geq 1$, $(\mu_t^{M,N})_{t\geq 0}$ be a family of laws on Θ^M ergodic to the empirical M-particle Gibbs measure (Assumption (E); for the M-particle empirical noisy gradient descent this is the ergodic theorem). Let g be bounded measurable, $\varepsilon > 0$ and $\delta \in (0,1)$. Then for the (λ, β) and N_1 of Theorem 427, on the event of probability at least $1-\delta$ over the sample, for every $M \geq 1$*

$$\limsup_{t\to\infty} \mathbb{E}\Big|\int g \,\mathrm{d}\Pi\rho_t^{M,N} - \int g\, u_0^\dagger \,\mathrm{d}\nu_0\Big| \leq \frac{C_g^N}{\sqrt{M}} + \frac{2\varepsilon}{3},$$

in the form: for every $\varepsilon' > 0$, eventually in t, $\mathbb{E}_{\mu_t}|\int g\,\mathrm{d}\Pi\rho_\theta - \int g\,u_0^\dagger\,\mathrm{d}\nu_0| \leq \frac{2\varepsilon}{3} + \frac{C_g(Y_{\max},\lambda,\beta)}{\sqrt{M}} + \varepsilon'$, with the constant of Definition 135. (The triangle inequality (**??**) *at fixed (M,t), integrated over μ_t: the second and third terms are deterministic on the event of stage (2) and bounded by $\varepsilon/3$ each, and the first is Theorem 554.)*

Proof. Stage (1) at '(λ, β)'.
Stage (2) at '(λ, β)', on the event 'E'_δ'.
Stage (3) for the sample 'ω' (labels bounded by 'Ymax' on 'E'_δ').
The triangle inequality 'eq:three-stages', integrated over 'μ_t'. ☐

Theorem 556. *Under the hypotheses of Corollary 555, since $C_g(Y_{\max}, \lambda, \beta)$ does not depend on the sample, there is $M_0 = M_0(\varepsilon; Y_{\max}, \lambda, \beta, g)$ such that on the same event of probability at least $1-\delta$, for every $M \geq M_0$, eventually in t, $\mathbb{E}_{\mu_t}|\int g\,\mathrm{d}\Pi\rho_\theta - \int g\,u_0^\dagger\,\mathrm{d}\nu_0| \leq \varepsilon$; that is,*

$$\limsup_{M\to\infty}\limsup_{t\to\infty} \mathbb{E}\Big|\int g\,\mathrm{d}\Pi\rho_t^{M,N} - \int g\,u_0^\dagger\,\mathrm{d}\nu_0\Big| \leq \frac{2\varepsilon}{3} < \varepsilon.$$

($M_0 := \max\{\lceil(6C_g/\varepsilon)^2\rceil, 1\}$ gives $C_g/\sqrt{M} \leq \varepsilon/6$; take $\varepsilon' = \varepsilon/6$.)

Proof. ☐

Chapter 9

Explicit geometry of the limit

Section 11 of the paper: for inputs uniform on the circle and a Gaussian hidden law the kernel K_{ν_0} is zonal and is diagonalized by the trigonometric system; representability, the canonical ridgelet transform and the source condition are read off from the Fourier coefficients.

Definition 557. Let $P_X = \tau$ be the normalized arc-length measure on $S^1 = \mathbb{R}/2\pi\mathbb{Z}$, i.e. the Haar probability measure of the additive circle.

Definition 558. For $\theta \in \mathbb{R}/2\pi\mathbb{Z}$ let $x(\theta) := (\cos\theta, \sin\theta) \in \mathbb{R}^2$, identified with $e^{i\theta} \in \mathbb{C} = \mathbb{R}^2$.

Definition 559. Let $\sigma\colon \mathbb{R} \to \mathbb{R}$ be measurable with $|\sigma| \le 1$ (tanh or erf). The feature on S^1 with hidden parameter $z = (w, b) \in \mathbb{R}^2 \times \mathbb{R}$ is $\varphi_z(x(\theta)) := \sigma(w \cdot x(\theta) - b) = \sigma(w_1 \cos\theta + w_2 \sin\theta - b)$, a feature map in the sense of Definition 1.

Definition 560. $\mathcal{N}(0, \sigma_w^2 I_2)$ on $\mathbb{R}^2$, the law of $\sigma_w G$ for a standard Gaussian vector G of $\mathbb{R}^2$.

Definition 561. The hidden law is $\nu_0 = \mathcal{N}(0, \sigma_w^2 I_2) \otimes \mathcal{N}(0, \sigma_b^2)$ on $Z = \mathbb{R}^2 \times \mathbb{R}$, $\sigma_w > 0$, $\sigma_b \ge 0$ (for $\sigma_b = 0$ the bias law is δ_0).

Definition 562. For a feature map φ and $\nu \in \mathcal{P}(Z)$ the *kernel function* is $k_\nu(x, x') := \int_Z \varphi_z(x)\varphi_z(x')\,\nu(\mathrm{d}z)$.

Definition 563. The *zonal function* of the kernel is $\kappa(\phi) := k_{\nu_0}(x(\phi), x(0))$, $\phi \in \mathbb{R}/2\pi\mathbb{Z}$; by rotation invariance of ν_0, $k_{\nu_0}(x(\theta), x(\theta')) = \kappa(\theta - \theta')$, and $\kappa(\phi) = \tilde{\kappa}(\cos\phi)$ in the notation of Theorem **??**.

Definition 564. The real Fourier modes $\cos(\ell\theta)$, $\sin(\ell\theta)$ ($\ell \in \mathbb{Z}$) as elements of $L^2(\tau)$; $\mathcal{H}_0 = \mathrm{span}\{1\}$ and $\mathcal{H}_\ell = \mathrm{span}\{\cos\ell\theta, \sin\ell\theta\}$ for $\ell \ge 1$.

Definition 565. The real trigonometric system $(e_\ell)_{\ell\in\mathbb{Z}}$ of $L^2(\tau)$: $e_0 := 1$, $e_\ell := \sqrt{2}\cos(\ell\theta)$ for $\ell > 0$ and $e_\ell := \sqrt{2}\sin(\ell\theta)$ for $\ell < 0$. It is an orthonormal basis of $L^2(\tau)$, $\{e_\ell, e_{-\ell}\}$ is an orthonormal basis of $\mathcal{H}_\ell$ ($\ell \ge 1$), and $\|P_\ell f\|^2 = \langle f, e_\ell\rangle^2 + \langle f, e_{-\ell}\rangle^2$.

Definition 566. $\mu_\ell := \frac{1}{2\pi}\int_0^{2\pi} \tilde{\kappa}(\cos\phi)\cos(\ell\phi)\,\mathrm{d}\phi = \int \kappa(\phi)\cos(\ell\phi)\,\tau(\mathrm{d}\phi)$, $\ell \in \mathbb{Z}$ (even in ℓ).

Definition 567. $a_k(s) := \mathbb{E}_{G\sim\gamma}[\sigma(sG)h_k(G)]$ for $k \ge 0$, $s \ge 0$, where $\gamma = \mathcal{N}(0,1)$ and $h_k = He_k/\sqrt{k!}$ are the normalized probabilists' Hermite polynomials.

Definition 568. $\kappa_s(\rho) := \sum_{k\ge 0} a_k(s)^2\rho^k$, $\rho \in [-1, 1]$ (the dual activation of Daniely et al.); the series converges absolutely since $\sum_k a_k(s)^2 = \|\sigma(s\cdot)\|^2_{L^2(\gamma)} \le \|\sigma\|^2_\infty$.

Definition 569. $\operatorname{erf}(u) := \frac{2}{\sqrt{\pi}} \int_0^u e^{-t^2}\,\mathrm{d}t$; it is odd, continuous, real analytic and $|\operatorname{erf}| \le 1$.

Theorem 570. *For $\|x\| = \|x'\| = 1$ and $s^2 = \sigma_w^2 + \sigma_b^2 > 0$, the pair $U = (w \cdot x - b)/s$, $U' = (w \cdot x' - b)/s$ is, under $\nu_0 = \mathcal{N}(0, \sigma_w^2 I_2) \otimes \mathcal{N}(0, \sigma_b^2)$, a standard Gaussian pair with correlation $\rho = (\sigma_w^2\, x \cdot x' + \sigma_b^2)/s^2$. (Both are centered Gaussian vectors with unit variances and covariance ρ; compare the characteristic functions.)*

Proof. '|ρ| ≤ 1' by Cauchy–Schwarz.
The left-hand side: 'exp(−(σ_w² ‖(αx + βx')/s‖² + σ_b² ((α + β)/s)²)/2)'.
The right-hand side: 'exp(−((α + βρ)² + β²(1 − ρ²))/2)'.
The covariances agree: 'σ_w² ‖αx + βx'‖²/s² + σ_b² (α + β)²/s² = α² + 2αβρ + β²'. ☐

Theorem 571. *For $s^2 = \sigma_w^2 + \sigma_b^2 > 0$ and $\theta, \theta' \in S^1$, $k_{\nu_0}(x(\theta), x(\theta')) = \sum_{k \ge 0} a_k(s)^2 \rho^k = \kappa_s(\rho)$ with $\rho = (\sigma_w^2\, x(\theta) \cdot x(\theta') + \sigma_b^2)/s^2$: the pre-activations $U = (w \cdot x(\theta) - b)/s$, $U' = (w \cdot x(\theta') - b)/s$ form a standard Gaussian pair with correlation ρ (Lemma 570) and $\mathbb{E}[\phi(U)\phi(U')] = \sum_k a_k(\phi)^2 \rho^k$ for $\phi = \sigma(s\cdot) \in L^2(\gamma)$ (Lemma 919).*

Proof. ☐

Theorem 572. *For $f \in C^1(\mathbb{R})$ with f and f' bounded and every $n \ge 0$, $\int f\, He_{n+1}\,\mathrm{d}\gamma = \int f'\, He_n\,\mathrm{d}\gamma$: since $He_{n+1} = XHe_n - He_n'$ and $(He_n \gamma)' = (He_n' - XHe_n)\gamma$, this is the integration by parts $\int f\,(XHe_n - He_n')\gamma = \int f' He_n \gamma$ (the boundary terms vanish by the Gaussian decay).*

Proof. Integrability of 'f P φ' and 'f' P φ' for polynomials 'P'.
The integration by parts on the line: '∫ u v' = −∫ u' v'.
Translate back to Gaussian integrals: '∫ f X He_n dγ = ∫ f' He_n dγ + ∫ f He_n' dγ'. ☐

Theorem 573. *For $b \ge 0$ and $v = (1 + 2b)^{-1}$, $\int He_n(y) e^{-by^2}\,\gamma(\mathrm{d}y) = \sqrt{v}\,(1 - v)^{n/2} He_n(0)$: the density $\gamma(y)e^{-by^2}$ is $\sqrt{v}$ times the density of $\mathcal{N}(0, v)$, and $\mathbb{E}[He_n(\sqrt{v}G)] = (1 - v)^{n/2} He_n(0)$ by the Gaussian smoothing identity.*

Proof. ☐

Theorem 574. $\frac{1}{2\pi} \int_0^{2\pi} \cos^n \phi \cos(\ell\phi)\,\mathrm{d}\phi = 2^{-n}\binom{n}{(n-|\ell|)/2}$ *for $|\ell| \le n$, $n \equiv \ell \pmod 2$, and 0 otherwise (binomial expansion of $\cos^n \phi = 2^{-n}(e^{i\phi} + e^{-i\phi})^n$ and orthogonality).*

Proof. ☐

Theorem 575. *For $s^2 = \sigma_w^2 + \sigma_b^2 > 0$, $\mu_\ell = \sum_{k \ge 0} a_k(s)^2 c_{k,\ell}$ with $c_{k,\ell} := \int \rho(\phi)^k \cos(\ell\phi)\,\tau(\mathrm{d}\phi)$, $\rho(\phi) = (\sigma_w^2 \cos\phi + \sigma_b^2)/s^2$ (dominated convergence in $\mu_\ell = \int \kappa(\phi) \cos(\ell\phi)\,\tau(\mathrm{d}\phi)$ with $\kappa(\phi) = \sum_k a_k(s)^2 \rho(\phi)^k$, the terms being bounded by the summable $a_k(s)^2$).*

Proof. ☐

Theorem 576. *If $f \in L^2(\gamma)$ is not γ-a.e. equal to a polynomial, then $a_k(f) \ne 0$ for infinitely many k: a finite Hermite expansion is a polynomial, by the completeness of the Hermite polynomials.*

Proof. ☐

Theorem 577. *For $r \in L^2(P_X)$, P_X-a.e. in x, $(K_\nu r)(x) = \int_{\mathcal{X}} k_\nu(x, x') r(x')\,P_X(\mathrm{d}x')$ with $k_\nu(x, x') = \int \varphi_z(x)\varphi_z(x')\,\nu(\mathrm{d}z)$ (Fubini in Lemma 194; the integrand is bounded by $|r(x')|$).*

Proof. The integrand '(z, x') ↦ φ_z(x) φ_z(x') r(x')' is integrable on 'ν ⊗ P'. ☐

Theorem 578. *The hidden law* $\nu_0 = \mathcal{N}(0, \sigma_w^2 I_2) \otimes \mathcal{N}(0, \sigma_b^2)$ *is invariant under the rotations* $(w, b) \mapsto (Rw, b)$ *of the weight,* $R \in SO(2)$*.*

Proof. $\square$

Theorem 579. $k_{\nu_0}(x(\theta), x(\theta')) = \kappa(\theta - \theta')$ *with the zonal function* κ *of Definition 563: rotating the weight* w *by the angle* θ' *leaves* ν_0 *invariant and moves* $x(\theta), x(\theta')$ *to* $x(\theta - \theta'), x(0)$*.*

Proof. $\square$

Theorem 580. K_{ν_0} *is the convolution operator* $g \mapsto \int \kappa(\theta - \theta') g(\theta')\, \tau(\,\mathrm{d}\theta')$ *on* $L^2(\tau)$*, so* $\cos(\ell\theta)$ *and* $\sin(\ell\theta)$ *are eigenfunctions with eigenvalue* $\mu_\ell = \frac{1}{2\pi} \int_0^{2\pi} \tilde{\kappa}(\cos\phi) \cos(\ell\phi)\, \mathrm{d}\phi$*, the* ℓ*th Fourier (cosine, by evenness) coefficient of* κ*:* $K_{\nu_0} = \sum_\ell \mu_\ell P_\ell$*.*

Proof. $\square$

Theorem 581. *The real trigonometric system* $(e_\ell)_{\ell \in \mathbb{Z}}$ *of Definition 565 is orthonormal in* $L^2(\tau)$*.*

Proof. $\square$

Definition 582. The real trigonometric system $(e_\ell)_{\ell \in \mathbb{Z}}$ is an orthonormal basis of $L^2(\tau)$: a function orthogonal to all $\cos(\ell\theta), \sin(\ell\theta)$ has vanishing complex Fourier coefficients, hence vanishes.

Definition 583. Let $(e_i)_{i \in I}$ be an orthonormal basis of $L^2(P_X)$ with $K_\nu e_i = \mu_i e_i$. The *singular vectors* of S_ν are $v_i := \mu_i^{-1/2} S_\nu^* e_i$ for $\mu_i \neq 0$.

Theorem 584. *Let* $(e_i)_{i \in I}$ *be an orthonormal basis of* $L^2(P_X)$ *with* $K_\nu e_i = \mu_i e_i$ *(so* $\mu_i = \|S_\nu^* e_i\|^2 \in [0, 1]$*). Then* $f \in \operatorname{ran} S_\nu \iff \sum_i \mu_i^{-1} \langle f, e_i \rangle^2 < \infty$ *and* $\langle f, e_i \rangle = 0$ *whenever* $\mu_i = 0$*. In that case* $u_0^\dagger = S_\nu^\dagger f = \sum_i \mu_i^{-1} \langle f, e_i \rangle S_\nu^* e_i$*, the series converging in* $L^2(\nu)$ *with mutually orthogonal terms, and* $\|u_0^\dagger\|_{L^2(\nu)}^2 = \sum_i \mu_i^{-1} \langle f, e_i \rangle^2$*. (Singular system:* $v_i := \mu_i^{-1/2} S_\nu^* e_i$ *is orthonormal with* $S_\nu v_i = \mu_i^{1/2} e_i$*; "*$\Rightarrow$*" is Bessel's inequality for* $\mu_i^{-1/2} \langle S_\nu u, e_i \rangle = \langle u, v_i \rangle$*, "*$\Leftarrow$*" is the termwise synthesis of* $u = \sum_i \mu_i^{-1/2} \langle f, e_i \rangle v_i$*, which lies in* $\overline{\operatorname{ran} S_\nu^*} = (\ker S_\nu)^\perp$*.)*

Proof. $\square$

Theorem 585. $0 \le \mu_\ell \le 1$ *for every* ℓ*:* $\mu_\ell = \|S_{\nu_0}^* e_\ell\|^2$ *and* $\|S_{\nu_0}^*\| \le 1$*.*

Proof. $\square$

Theorem 586. *Assume* $\mu_\ell > 0$ *for all* ℓ *(by (ii) this holds when* $\sigma_b > 0$ *and* σ *is not a polynomial). Then* $f \in \operatorname{ran} S_{\nu_0} \iff \sum_{\ell \ge 0} \mu_\ell^{-1} \|P_\ell f\|_{L^2(\tau)}^2 < \infty$*, where* $\|P_\ell f\|^2 = \langle f, e_\ell \rangle^2 + \langle f, e_{-\ell} \rangle^2$ *in the real trigonometric basis* $(e_\ell)_{\ell \in \mathbb{Z}}$ *of Definition 565, i.e.* $\sum_{\ell \in \mathbb{Z}} \mu_\ell^{-1} \langle f, e_\ell \rangle^2 < \infty$*. (Lemma 584 for the eigenbasis of Theorem 580.)*

Proof. $\square$

Theorem 587. *Under (R),* $u_0^\dagger = S_{\nu_0}^\dagger f = \sum_{\ell \in \mathbb{Z}} \mu_\ell^{-1} \langle f, e_\ell \rangle S^* e_\ell = \sum_{\ell \ge 0} \mu_\ell^{-1} S^* P_\ell f$*, with* $(S^* e_\ell)(w, b) = \frac{1}{2\pi} \int_0^{2\pi} \sigma(w_1 \cos\theta + w_2 \sin\theta - b) e_\ell(\theta)\, \mathrm{d}\theta$*, the series converging in* $L^2(\nu_0)$ *with mutually orthogonal terms, and* $\|u_0^\dagger\|_{L^2(\nu_0)}^2 = \sum_{\ell \in \mathbb{Z}} \mu_\ell^{-1} \langle f, e_\ell \rangle^2 = \sum_{\ell \ge 0} \mu_\ell^{-1} \|P_\ell f\|^2$*.*

Proof. $\square$

Theorem 588. *Under (R), Theorem* **??** *($\lambda_n \to 0$, $\kappa_n \to 0$) gives $m_n^* \to u_0^\dagger$ in $L^2(\nu_0)$ and $\Pi\rho_n^* \to u_0^\dagger\nu_0$ in total variation: the learned conditional mean amplitude converges to the explicit weighted ridgelet transform $u_0^\dagger = \sum_\ell \mu_\ell^{-1} S^* P_\ell f$ of Theorem 587, the τ-weighted ridgelet transform S^* with the activation as filter applied to the sharpened target $K_{\nu_0}^{-1} f = \sum_\ell \mu_\ell^{-1} P_\ell f$.*

Proof. □

Theorem 589. *For all $x, x' \in \mathbb{R}^m$, $k_{\nu_0}(x,x') = \int \varphi_z(x)\varphi_z(x')\nu_0(\mathrm{d}z) = \sum_{k\geq 0} a_k(s_x)a_k(s_{x'})\rho_{xx'}^k$, the series converging absolutely; on the circle $s_x \equiv s$ and $\kappa(\phi) = \tilde{\kappa}(\cos\phi) = \kappa_s((\sigma_w^2\cos\phi + \sigma_b^2)/s^2)$. (In Lean: the case $\|x\| = \|x'\| = 1$.)*

Proof. The pre-activations at ‘x(φ)‘ and ‘x(0)‘ form a standard Gaussian pair with correlation ‘ρ(φ) = (σ_w² cos φ + σ_b²)/s²‘, and the Mehler-type identity gives ‘κ(φ) = ∑_k a_k(s)² ρ(φ)^k = κ_s(ρ(φ))‘. □

Theorem 590. *$\mu_\ell = \sum_{n\geq\ell,\ n\equiv\ell\ (2)} c_n 2^{-n}\binom{n}{(n-\ell)/2}$ with $c_n := \sum_{k\geq n} a_k(s)^2\binom{k}{n}(\sigma_w^2/s^2)^n(\sigma_b^2/s^2)^{k-n}$: expand $\tilde{\kappa}(t) = \sum_n c_n t^n$ (nonnegative coefficients, Tonelli) and insert $\frac{1}{2\pi}\int_0^{2\pi}\cos^n\phi\cos(\ell\phi)\,\mathrm{d}\phi = 2^{-n}\binom{n}{(n-\ell)/2}$.*

Proof. ‘μ_ℓ = ∑_k a_k² c_k,ℓ‘ with ‘c_k,ℓ = ∫ ρ^k cos(ℓ·) dτ‘, expand ‘ρ^k‘ binomially and rearrange the double series of nonnegative terms (‘mercerEigenvalue_eq_tsum_zonalPowerCoeff‘), then insert the Fourier coefficients of ‘cos^n‘ (‘lem:cos-pow-fourier‘). □

Theorem 591. *If $\sigma_b > 0$ and σ is not (a.e. equal to) a polynomial, then $\mu_\ell > 0$ for all ℓ, K_{ν_0} is injective and $\overline{\operatorname{ran} S_{\nu_0}} = L^2(\tau)$. (A non-polynomial σ has $a_k(s) \neq 0$ for infinitely many k, so every $c_n > 0$ and $\mu_\ell \geq c_\ell 2^{-\ell} > 0$.)*

Proof. Step 1: ‘σ(s·)‘ is not ‘γ‘-a.e. a polynomial, so ‘a_k(s) ≠ 0‘ for some ‘k ≥ |ℓ|‘ (completeness of the Hermite polynomials).

Step 2: ‘μ_ℓ = ∑_k a_k² c_k,ℓ ≥ a_k² c_k,ℓ > 0‘, since ‘c_k,ℓ > 0‘ for ‘k ≥ |ℓ|‘ when ‘σ_b > 0‘. □

Theorem 592. *Assume $\mu_\ell > 0$ for all ℓ. For $a \in \{1/2, 1\}$, $u_0^\dagger \in \operatorname{ran}((S_{\nu_0}^* S_{\nu_0})^a) \iff \sum_{\ell\in\mathbb{Z}} \mu_\ell^{-2a-1}\langle f, e_\ell\rangle^2 < \infty$ $(= \sum_{\ell\geq 0}\mu_\ell^{-2a-1}\|P_\ell f\|^2)$, the source condition (SC$_a$) of Definition 139 for some source norm G. (In the singular system of Lemma 584: $a = 1/2$ is $f \in \operatorname{ran} K_{\nu_0}$, $a = 1$ is $f \in K_{\nu_0}(\operatorname{ran} S_{\nu_0})$, both read off the coordinates $\langle e_\ell, K_{\nu_0} g\rangle = \mu_\ell\langle e_\ell, g\rangle$.)*

Proof. □

Theorem 593. *Assume $\mu_\ell > 0$ for all ℓ. For every $a > 0$, $u_0^\dagger \in \operatorname{ran}((S_{\nu_0}^* S_{\nu_0})^a) \iff \sum_{\ell\in\mathbb{Z}} \mu_\ell^{-2a-1}\langle f, e_\ell\rangle^2 < \infty$, where T_0^a is the spectral power of Definition 164. (In the singular system, $T_0 v_\ell = \mu_\ell v_\ell$ and $(T_0)^a$ maps $v_\ell \mapsto \mu_\ell^a v_\ell$ on $(\ker S_{\nu_0})^\perp$ and vanishes on $\ker S_{\nu_0}$.)*

Proof. □

Theorem 594. *If $\sigma_b = 0$ and σ is odd, then $\mu_\ell = 0$ for even ℓ, and $\operatorname{ran} S_{\nu_0}$ consists of odd functions. ($\kappa(\phi+\pi) = -\kappa(\phi)$, so $\mu_\ell = \int \kappa(\phi+\pi)\cos(\ell(\phi+\pi))\,\tau(\mathrm{d}\phi) = -\mu_\ell$ for even ℓ.)*

Proof. □

Theorem 595. *For $\sigma = \operatorname{erf}$: $a_{2j}(s) = 0$ and $a_{2j+1}(s) = \frac{2}{\sqrt{\pi}}\frac{(-1)^j(2j)!}{j!\sqrt{(2j+1)!}}\frac{s^{2j+1}}{(1+2s^2)^{j+1/2}}$ $(j \geq 0)$; in particular $a_1(s) = \frac{2}{\sqrt{\pi}}\frac{s}{\sqrt{1+2s^2}}$. (Gaussian integration by parts, $\sqrt{k!}\,a_k(s) = s^k\mathbb{E}[\sigma^{(k)}(sG)]$, and the generating function of the physicists' Hermite polynomials.)*

Proof. □

Theorem 596. *For $\sigma = \mathrm{erf}$, $\kappa_s(\rho) = \frac{2}{\pi}\arcsin\frac{2s^2\rho}{1+2s^2}$, hence on the circle $\kappa(\phi) = \tilde{\kappa}(\cos\phi) = \frac{2}{\pi}\arcsin\frac{2(\sigma_w^2\cos\phi+\sigma_b^2)}{1+2s^2}$, $s^2 = \sigma_w^2 + \sigma_b^2$ (the arcsine kernel of Williams 1998; the series of Lemma 595 is the Taylor series of the arcsine).*

Proof. □

Theorem 597. *For $\sigma = \mathrm{erf}$ and $\sigma_b > 0$ there are $\eta > 0$ and $C < \infty$ with $\mu_\ell \le Ce^{-\eta|\ell|}$: $\phi \mapsto \tilde{\kappa}(\cos\phi)$ is 2π-periodic and holomorphic on a strip $|\Im\phi| < \eta$, since $|2(\sigma_w^2 t + \sigma_b^2)/(1+2s^2)| < 1$ on $[-1,1]$, and its Fourier coefficients decay geometrically.*

Proof. ‘μ_ℓ ≤ ∑_k ≥ |ℓ| a_k(s)^2 ≤ (2/π) ∑_k ≥ |ℓ| q^k = (2/(π(1 − q))) q^|ℓ|‘ with ‘q = 2s^2/(1 + 2s^2) < 1‘, from the explicit Hermite coefficients of erf (‘mercerEigenvalue_erf_le_pow‘); take ‘η = −log q‘. □

Theorem 598. *For $\sigma = \mathrm{erf}$, $\sigma_b > 0$, (R) implies $\sum_\ell e^{\eta|\ell|}\|P_\ell f\|^2 < \infty$ for some $\eta > 0$, so f is real analytic in θ (holomorphic extension to a strip $|\Im\theta| < \eta/2$): $\|P_\ell f\|^2 \le \|u_0^\dagger\|^2\mu_\ell \le C\|u_0^\dagger\|^2 e^{-\eta|\ell|}$.*

Proof. □

Theorem 599. *For $\sigma = \tanh$ and $\sigma_b > 0$, $\mu_\ell = O(|\ell|^{-p})$ for every $p > 0$: $\sum_k k^{2p} a_k(s)^2 = \|N^p \tanh(s\cdot)\|^2_{L^2(\gamma)} < \infty$ for the Ornstein–Uhlenbeck number operator N, so $\phi \mapsto \tilde{\kappa}(\cos\phi)$ is C^∞ and its Fourier coefficients decay faster than any power.*

Proof. ‘f = tanh(s·)‘ is smooth with bounded derivatives ‘s^i tanh^(i)(s·)‘.

With ‘m = ⌈p⌉_+‘, ‘μ_ℓ ≤ ‖f^(m)‖^2 / (|ℓ| − m + 1)^m‘ for ‘|ℓ| ≥ m‘ (‘mercerEigenvalue_le_of_smooth‘, by iterated Gaussian integration by parts).

Large ‘|ℓ|‘: ‘(|ℓ| − m + 1)^-m ≤ 2^m |ℓ|^-m ≤ 2^m |ℓ|^-p‘.

Small ‘|ℓ| < 2m‘: ‘μ_ℓ ≤ 1 ≤ C |ℓ|^-p‘. □

Theorem 600. *For $\sigma = \tanh$, $\sigma_b > 0$, (R) implies $\sum_\ell |\ell|^p\|P_\ell f\|^2 < \infty$ for every p, so $\|P_\ell f\| = O(|\ell|^{-p})$ for every p and $f \in C^\infty(S^1)$.*

Proof. □

Chapter 10

Heavy-tailed hidden laws and Sobolev-type representability

Section 12 of the paper: for $m = 1$ and $\sigma = \tanh$, the truncated homogeneous hidden law $\nu_0^{(L)}$ with density $|w|^{\alpha-1}$ on a box, whose kernel converges to the sign-network kernel $K^{(\infty)}(x, x') = 1 - c_\alpha |x - x'|$ as $L \to \infty$; representability becomes $f \in H^1$ in the limit.

Definition 601. For a feature map φ and $\nu \in \mathcal{P}(Z)$, the *kernel function* is $K_\nu(x, x') := \int \varphi_z(x)\varphi_z(x')\,\nu(\mathrm{d}z)$, the integral kernel of $K_\nu = S_\nu S_\nu^*$ (Definition 18).

Definition 602. (Conventions of Section 10.) $m = 1$, the hidden parameter is $z = (w, b) \in Z = \mathbb{R}^2$ and $\varphi_z(x) = \tanh(wx - b)$; this is a feature map in the sense of Definition 1.

Theorem 603. *For $m = 1$ the feature $\varphi_{(w,b)}(x) = \tanh(w^\top x - b)$ of Definition 5 is $\tanh(w_0 x_0 - b)$ under $\mathbb{R}^1 \cong \mathbb{R}$.*

Proof. □

Definition 604. $C_\alpha := \frac{1}{1-2^{-2\alpha}}\left(2 + \frac{2\alpha}{1+\alpha} + \frac{4\alpha}{3(2+\alpha)} + 2^{1-\alpha}\alpha\right) + 1$ (**??**).

Definition 605. $\ell_\alpha := \frac{1}{c_\alpha} = \frac{1+\alpha}{\alpha}$ (**??**).

Definition 606. $M_L := \int_{1/L}^{L} w^{\alpha-1}\,\mathrm{d}w = \frac{L^\alpha - L^{-\alpha}}{\alpha}$ (Lemma 649).

Definition 607. $Z_L := 4LM_L = \frac{4L(L^\alpha - L^{-\alpha})}{\alpha}$, the value of $\int_{B_L} |w|^{\alpha-1}\,\mathrm{d}w\,\mathrm{d}b$ (Lemma 649).

Definition 608. The annulus $A_L := \{w \in \mathbb{R} : 1/L \le |w| \le L\}$ in the scale variable, so that $B_L = A_L \times [-L, L]$.

Definition 609. $B_L := \{(w, b) : 1/L \le |w| \le L, \ |b| \le L\}$ (**??**).

Definition 610. Let $0 < \alpha < 1$ and $L \ge 2$. The *truncated homogeneous law* is $\nu_0^{(L)}(\mathrm{d}w\,\mathrm{d}b) := Z_L^{-1}|w|^{\alpha-1}\mathbf{1}_{B_L}(w, b)\,\mathrm{d}w\,\mathrm{d}b$ with $B_L = \{(w, b) : 1/L \le |w| \le L, \ |b| \le L\}$ and $Z_L = \int_{B_L} |w|^{\alpha-1}\,\mathrm{d}w\,\mathrm{d}b$ (**??**).

Definition 611. $K^{(L)}(x, x') := K_{\nu_0^{(L)}}(x, x') = \int \tanh(wx - b)\tanh(wx' - b)\,\nu_0^{(L)}(\mathrm{d}w\,\mathrm{d}b)$ (Definition 610).

Definition 612. The *sign feature* with threshold $t \in \mathbb{R}$ is $\varphi_t(x) := \operatorname{sgn}(x - t)$ ($= 0$ at $x = t$), a feature map on $Z = \mathcal{X} = \mathbb{R}$ (Remark 655).

Definition 613. $\varpi := U[-\ell_\alpha, \ell_\alpha]$, the uniform law $\frac{\mathrm{d}t}{2\ell_\alpha}$ on $[-\ell_\alpha, \ell_\alpha]$ (Remark 655).

Definition 614. (Conventions of Section 10.) $\mathcal{X} = [-1, 1]$ and $P_X(\mathrm{d}x) = \frac{1}{2}\,\mathrm{d}x$ on $[-1, 1]$, so that (A1) holds with $R_X = 1$.

Definition 615. $K^{(\infty)}(x, x') := 1 - c_\alpha|x - x'| = 1 - \frac{\alpha}{1+\alpha}|x - x'|$ (Theorem 676).

Definition 616. $s \coth s = \frac{s \cosh s}{\sinh s}$ for $s \neq 0$, extended by its limit 1 at $s = 0$ (the value of the bias integral (**??**) divided by 2).

Definition 617. $I_\alpha(S) := \int_0^S (s \coth s - 1) s^{\alpha-1}\,\mathrm{d}s$ for $S > 0$ (**??**).

Definition 618. $J_\alpha(S) := \int_0^S \frac{2s^\alpha}{e^{2s}-1}\,\mathrm{d}s$ (**??**).

Definition 619. $D_\alpha := \int_0^\infty \frac{2s^\alpha}{e^{2s}-1}\,\mathrm{d}s = 2^{-\alpha}\Gamma(1+\alpha)\zeta(1+\alpha)$ (**??**).

Definition 620. $\Psi_L(d) := \int_{1/L}^L wd \coth(wd)\, w^{\alpha-1}\,\mathrm{d}w$ for $d \neq 0$ and $\Psi_L(0) := M_L$ (**??**).

Definition 621. $E_L(w; x, x') := \int_{|b|>L} [1 - \tanh(wx - b)\tanh(wx' - b)]\,\mathrm{d}b$ (**??**).

Definition 622. $\mathcal{E}_L(x, x') := \int_{1/L \le |w| \le L} E_L(w; x, x')\,|w|^{\alpha-1}\,\mathrm{d}w$ (**??**).

Definition 623. $T_1 := \frac{L^{-2\alpha}}{L(1-L^{-2\alpha})}$ (Theorem 676).

Definition 624. $T_2 := \frac{\alpha L^{-2\alpha}}{(1+\alpha)(1-L^{-2\alpha})}|d|$ (Theorem 676).

Definition 625. $T_3 := \frac{|d|^{-\alpha} J_\alpha(L|d|)}{LM_L}$ (Theorem 676); in Lean $0^{-\alpha} = 0$, so $T_3 = 0$ at $d = 0$ (the paper takes the limit $\frac{1}{L} + T_1$ there).

Definition 626. $T_4 := \frac{|d|^{-\alpha} I_\alpha(|d|/L)}{LM_L}$ (Theorem 676).

Definition 627. $T_5 := \frac{\mathcal{E}_L(x,x')}{4LM_L}$ (Theorem 676).

Definition 628. $C_\alpha := \frac{1}{1-2^{-2\alpha}}\left(2 + \frac{2\alpha}{1+\alpha} + \frac{4\alpha}{3(2+\alpha)} + 2^{1-\alpha}\alpha\right) + 1$ (**??**).

Definition 629. For $L \in \mathbb{R}$ put $\nu_0^{[L]} := \nu_0^{(\max(L,2))}$ (Definition 610); this is $\nu_0^{(L)}$ for $L \ge 2$ and a probability measure for every L, so that the limits $L \to \infty$ can be taken along all real (or natural) L.

Definition 630. For $\lambda > 0$ and $f \in L^2(P_X)$, $R_\lambda^{(\infty)} f := S^*(K^{(\infty)} + \lambda)^{-1} f$, i.e. $(R_\lambda^{(\infty)} f)(w, b) = \int \tanh(wx - b)\,[(K^{(\infty)} + \lambda)^{-1} f](x)\,P_X(\mathrm{d}x)$, where S^* is the analysis map of the tanh feature and $K^{(\infty)}$ the kernel operator of the sign network (**??**).

Definition 631. For $f = K^{(\infty)} g \in \operatorname{ran} K^{(\infty)}$ (with $g \in L^2(P_X)$ unique by injectivity), $u_\infty^\dagger := S^*(K^{(\infty)})^{-1} f = S^* g$, i.e. $u_\infty^\dagger(w, b) = \int \tanh(wx - b) g(x)\,P_X(\mathrm{d}x)$ (**??**).

Definition 632. The integral operator of the limit kernel on functions on $[-1, 1]$, $K^{(\infty)} g(x) := \int_{-1}^1 K^{(\infty)}(x, x') g(x')\,P_X(\mathrm{d}x') = \frac{1}{2}\int_{-1}^1 (1 - c_\alpha|x - x'|) g(x')\,\mathrm{d}x'$ (Theorem **??**).

Definition 633. $Tg(x) := \int_{-1}^{1} |x-y|g(y)\,\mathrm{d}y$, so that $K^{(\infty)}g = \frac{1}{2}\langle g, 1\rangle_{L^2(\mathrm{d}x)} - \frac{c_\alpha}{2}Tg$ (proof of Theorem **??**).

Definition 634. $(K^{(\infty)}g)'(x) = -\frac{c_\alpha}{2}\left(\int_{-1}^{x} g - \int_{x}^{1} g\right)$, the derivative of $K^{(\infty)}g = \frac{1}{2}\int g - \frac{c_\alpha}{2}Tg$ with $(Tg)'(x) = \int_{-1}^{x} g - \int_{x}^{1} g$ (proof of Theorem **??**).

Definition 635. $\varkappa_j^e$, the unique root of $\varkappa \tan \varkappa = \alpha$ in $(j\pi, j\pi + \frac{\pi}{2})$ (Theorem **??**(iii)); $\mathcal{K}^e = \{\varkappa_j^e\}_{j\geq 0}$.

Definition 636. $\varkappa_{(2j)} := \varkappa_j^e$ and $\varkappa_{(2j+1)} := (j+\frac{1}{2})\pi$, the increasing enumeration of $\mathcal{K}^e \cup \mathcal{K}^o$ (Theorem **??**(iv)).

Definition 637. $\mu_k := \frac{c_\alpha}{\varkappa_{(k)}^2}$, the eigenvalues of $K^{(\infty)}$ in decreasing order (Theorem **??**(iv)).

Definition 638. $k_{\text{max}}(L) := \lfloor\sqrt{2c_\alpha/(\pi^2\varepsilon_L)}\rfloor - 1$ (Theorem **??**(b)).

Definition 639. For a bounded operator A on a real Hilbert space and a subspace V, $m_V(A) := \inf_{g\in V,\ \|g\|=1}\langle Ag, g\rangle$ (proof of Theorem **??**(a)).

Definition 640. $\mu_k(A) := \max_{\dim V = k+1} \min_{g\in V,\ \|g\|=1}\langle Ag, g\rangle$, the min–max value; for a compact self-adjoint positive operator it is the k-th eigenvalue in decreasing order (proof of Theorem **??**(a)).

Definition 641. For $\nu \in \mathcal{P}(Z)$, $\lambda > 0$ and $f \in L^2(P_X)$ the *resolvent quadratic form* is $Q_\lambda^\nu(f) := \langle f, (K_\nu + \lambda)^{-1} f\rangle_{L^2(P_X)}$, where $K_\nu = S_\nu S_\nu^*$ is the kernel operator and $(K_\nu+\lambda)^{-1}$ its resolvent (Definition 19). Equivalently, with $g = (K_\nu+\lambda)^{-1}f$, $Q_\lambda^\nu(f) = \|S^*g\|^2_{L^2(\nu)} + \lambda\|g\|^2_{L^2(P_X)}$; and $\frac{\lambda}{2}Q_\lambda^\nu(f) = \min_{u\in L^2(\nu)}[\frac{1}{2}\|S_\nu u - f\|^2 + \frac{\lambda}{2}\|u\|^2]$ (Proposition **??**).

Definition 642. $f \in H^1(-1,1)$ *with weak derivative* g if $g \in L^2(P_X)$ and $f(x) = f(-1) + \int_{-1}^{x} g(t)\,\mathrm{d}t$ for all $x \in [-1,1]$ (absolutely continuous with an L^2 derivative; $g = f'$ is unique a.e.). $H^1(-1,1)$ is the set of f admitting such a g.

Definition 643. $f \in H^2(-1,1)$ with $f' = g$ and $f'' = h$ if $f \in H^1(-1,1)$ with weak derivative g and $g \in H^1(-1,1)$ with weak derivative h.

Definition 644. For f with $f' = g$ continuous on $[-1,1]$, the *boundary conditions* of (**??**) are $f'(1) = -\frac{\alpha}{2}(f(1)+f(-1))$ and $f'(-1) = +\frac{\alpha}{2}(f(1)+f(-1))$.

Definition 645. For $f \in H^1(-1,1)$ with $f' = g$ and $c_0 := \frac{f(1)+f(-1)}{2}$, $\|f\|^2_{\mathcal{H}_\infty} := \frac{1+\alpha}{\alpha}\|f'\|^2_{L^2(P_X)} + (1+\alpha)c_0^2$ (**??**), where $\|f'\|^2_{L^2(P_X)} = \frac{1}{2}\int_{-1}^{1} f'(x)^2\,\mathrm{d}x$.

Definition 646. For $f \in H^1(-1,1)$ with $f' = g$ and $c_0 := \frac{f(1)+f(-1)}{2}$, $u_\varpi^\dagger(t) := \frac{1+\alpha}{\alpha}g(t)\mathbf{1}_{(-1,1)}(t) + (1+\alpha)c_0[\mathbf{1}_{(-\ell_\alpha,-1)}(t) - \mathbf{1}_{(1,\ell_\alpha)}(t)]$ (**??**).

Theorem 647. *For $s \neq 0$, $2s\coth s = 2|s| + \frac{4|s|}{e^{2|s|}-1}$ (**??**).*

Proof. □

Theorem 648. *For $\alpha > 0$ and $L \geq 1$, $\int_{1/L\leq|w|\leq L} |w|^{\alpha-1}\,\mathrm{d}w = 2M_L$ (Lemma 649).*

Proof. Step 1: split the annulus into the two intervals.
Step 2: the positive part is '∫_1/L^L w^α−1 dw = M_L'.
Step 3: the negative part equals the positive one by 'w ↦ −w'. □

Theorem 649. *With* $M_L := \int_{1/L}^{L} w^{\alpha-1}\,\mathrm{d}w = \frac{L^\alpha - L^{-\alpha}}{\alpha}$, $Z_L = \int_{B_L} |w|^{\alpha-1}\,\mathrm{d}w\,\mathrm{d}b = 2L\cdot 2M_L = 4LM_L = \frac{4L(L^\alpha-L^{-\alpha})}{\alpha}$ **(??)**.

Proof. Step 1: the box is a product and the integrand depends on ‘w‘ alone.
Step 2: the bias interval has length ‘2L‘. □

Theorem 650. *For* $\alpha > 0$ *and* $L > 1$, $\nu_0^{(L)}$ *is a probability measure:* $\int_{B_L} |w|^{\alpha-1}\,\mathrm{d}w\,\mathrm{d}b = Z_L$ *(Lemma 649).*

Proof. Step 1: the lower integral is the Bochner integral of the nonnegative density.
Step 2: ‘∫_B_L |w|^α−1 = Z_L‘. □

Theorem 651. *For* $\alpha > 0$ *and* $L > 1$, $\nu_0^{(L)}(\{|b| \le |w|\}) = \frac{\alpha}{1+\alpha}\cdot\frac{1-L^{-2-2\alpha}}{1-L^{-2\alpha}}$ **(??)***; it tends to* c_α *as* $L\to\infty$.

Proof. Step 1: the mass is the Bochner integral of the density over ‘|b| ≤ |w| ∩ B_L‘.
Step 2: Fubini: the inner integral in ‘b‘ over ‘[−|w|, |w|]‘ gives ‘2|w|^α‘ on the annulus.
Step 3: ‘∫_1/L ≤ |w| ≤ L 2|w|^α dw = 4 M_α+1,L‘. □

Theorem 652. *For* $u_1, u_2, b \in \mathbb{R}$ *and* $s := u_1 - u_2$, $1-\tanh(u_1-b)\tanh(u_2-b) = \frac{\cosh s}{\cosh(b-u_1)\cosh(b-u_2)}$ **(??)**.

Proof. □

Theorem 653. *For* $u_1, u_2 \in \mathbb{R}$ *with* $s := u_1 - u_2 \neq 0$, $\int_{\mathbb{R}} \frac{\mathrm{d}b}{\cosh(b-u_1)\cosh(b-u_2)} = \frac{2s}{\sinh s}$ **(??)**.

Proof. □

Theorem 654. *For* $u_1, u_2 \in \mathbb{R}$ *and* $s := u_1 - u_2$, $1-\tanh(u_1-b)\tanh(u_2-b) = \frac{\cosh s}{\cosh(b-u_1)\cosh(b-u_2)}$, $\int_{\mathbb{R}} \frac{\mathrm{d}b}{\cosh(b-u_1)\cosh(b-u_2)} = \frac{2s}{\sinh s}$, *hence* $\int_{\mathbb{R}}[1-\tanh(u_1-b)\tanh(u_2-b)]\,\mathrm{d}b = 2s\coth s = 2|s| + \frac{4|s|}{e^{2|s|}-1}$ *(*$s \neq 0$*;* $= 2$ *for* $s = 0$*).*

Proof. Step 1 (‘s = 0‘): the integrand is ‘1 − tanh²(b − u₁) = (tanh(b − u₁))′‘ and ‘tanh(±∞) = ±1‘.
Step 2 (‘s ≠ 0‘): ‘cosh s · 2s / sinh s = 2 s coth s‘. □

Theorem 655. *For* $\varpi := U[-\ell_\alpha, \ell_\alpha]$ *and* $x, x' \in [-1, 1]$, $\int_{-\ell_\alpha}^{\ell_\alpha} \operatorname{sgn}(x-t)\operatorname{sgn}(x'-t)\,\frac{\mathrm{d}t}{2\ell_\alpha} = 1 - \frac{|x-x'|}{\ell_\alpha} = K^{(\infty)}(x, x')$ **(??)**.

Proof. □

Theorem 656. *For* $|x|, |x'| \le 1$, $|w| \le L$ *and* $g_w(b) := \tanh(wx-b)\tanh(wx'-b)$, $E_L(w; x, x') := \int_{|b|>L}[1-g_w(b)]\,\mathrm{d}b \in [0,\ 4e^{-2(L-|w|)}]$ **(??)**.

Proof. Step 1: split ‘|b| > L‘ into the two half lines.
Step 2: the half line ‘b > L‘: ‘∫_L^∞ 4 e^2|w| e^−2b db = 2 e^2|w| e^−2L‘.
Step 3: the half line ‘b < −L‘: ‘∫_−∞^−L 4 e^2|w| e^2b db = 2 e^2|w| e^−2L‘.
Step 4: sum the two bounds. □

Theorem 657. *For* $S \ge 0$, $J_\alpha(S) \ge 0$.

Proof. □

Theorem 658. *For* $0 < \alpha$ *and* $S \ge 0$, $J_\alpha(S) \le S^\alpha/\alpha$ *(from* $\frac{2s}{e^{2s}-1} \le 1$*).*

Proof. □

Theorem 659. *For $0 < \alpha$ and $S \geq 0$, $J_\alpha(S) \leq D_\alpha = \int_0^\infty \frac{2s^\alpha}{e^{2s}-1}\,\mathrm{d}s < \infty$ (**??**).*

Proof. □

Theorem 660. *For $S \geq 0$, $I_\alpha(S) \geq 0$ (the integrand $(s \coth s - 1)s^{\alpha-1}$ is nonnegative).*

Proof. □

Theorem 661. *For $0 < \alpha$ and $S \geq 0$, $I_\alpha(S) \leq \frac{S^{2+\alpha}}{3(2+\alpha)}$ (from the Langevin bound $s \coth s - 1 \leq s^2/3$).*

Proof. □

Theorem 662. *For $0 < \alpha$ and $S \geq 0$, $I_\alpha(S) = \frac{S^{1+\alpha}}{1+\alpha} - \frac{S^\alpha}{\alpha} + J_\alpha(S)$ (**??**): integrate $(s \coth s - 1)s^{\alpha-1} = s^\alpha - s^{\alpha-1} + \frac{2s^\alpha}{e^{2s}-1}$, each term being integrable at 0.*

Proof. Step 1: the pointwise identity of the integrands on '[0, S]' (at 's = 0' both sides vanish).
Step 2: integrate term by term. □

Theorem 663. *For $L \geq 0$ and $g_w(b) = \tanh(wx - b)\tanh(wx' - b)$, $\int_{-L}^{L} g_w(b)\,\mathrm{d}b = 2L - \int_{\mathbb{R}}(1 - g_w)\,\mathrm{d}b + E_L(w; x, x') = 2L - 2wd\coth(wd) + E_L(w; x, x')$ with $d = x - x'$ ((**??**), (**??**)).*

Proof. Step 1: '∫_ℝ (1 − g_w) = ∫_|b| > L (1 − g_w) + ∫_[−L, L] (1 − g_w)'.
Step 2: '∫_[−L, L] g_w = 2L − ∫_[−L, L] (1 − g_w)'. □

Theorem 664. *$Z_L K^{(L)}(x, x') = \int_{B_L} \tanh(wx - b)\tanh(wx' - b)\,|w|^{\alpha-1}\,\mathrm{d}w\,\mathrm{d}b$ (Definition 610).*

Proof. □

Theorem 665. *For $0 < \alpha$, $L > 1$ and $x, x' \in [-1, 1]$, $d = x - x'$: $Z_L K^{(L)}(x, x') = 2L \cdot 2M_L - 4\Psi_L(d) + \mathcal{E}_L(x, x')$ (Fubini over B_L, the inner integral of Lemma 663, and the evenness of $wd\coth(wd)$ in w).*

Proof. Step 1: Fubini over 'B_L = A_L × [−L, L]'.
Step 2: the inner integral ('lem:ht-inner-bias-integral').
Step 3: split the outer integral into its three parts.
Step 4: '∫_A_L (wd) coth(wd) |w|^α−1 dw = 2 Ψ_L(d)' by evenness. □

Theorem 666. *For $0 < \alpha$, $L > 1$, $x, x' \in [-1, 1]$ and $d = x - x'$, $K^{(L)}(x, x') = 1 - \frac{\Psi_L(d)}{LM_L} + \frac{\mathcal{E}_L(x,x')}{4LM_L}$ (**??**).*

Proof. □

Theorem 667. *For $0 < \alpha < 1$, $L \geq 2$ and $x, x' \in [-1, 1]$, $0 \leq \mathcal{E}_L(x, x') \leq 2^{3-\alpha}L^{\alpha-1} + 8e^{-L}M_L$ (**??**): by (**??**), $\mathcal{E}_L \leq 8\int_{1/L}^{L} e^{-2(L-w)}w^{\alpha-1}\,\mathrm{d}w$; on $[1/L, L/2]$, $e^{-2(L-w)} \leq e^{-L}$, and on $[L/2, L]$, $w^{\alpha-1} \leq (L/2)^{\alpha-1}$ and $\int_{L/2}^{L} e^{-2(L-w)}\,\mathrm{d}w \leq \frac{1}{2}$.*

Proof. Step 1: 'E_L ≤ 4 e^−2(L − |w|)' pointwise ('lem:bias-truncation').
Step 2: evenness and '|w| = w' on '[1/L, L]'.
Step 3: split at 'L/2'.
Step 4: on '[1/L, L/2]', 'e^−2(L − w) ≤ e^−L' and '∫_1/L^L/2 w^α−1 ≤ M_L'.
Step 5: on '[L/2, L]', 'w^α−1 ≤ (L/2)^α−1' and '∫_L/2^L e^−2(L − w) ≤ 1/2'. □

Theorem 668. $\Psi_L(0) = M_L$ (**??**).

Proof. □

Theorem 669. *For $0 < \alpha$, $L \geq 1$ and $d \neq 0$, $\Psi_L(d) = \int_{1/L}^{L} wd\coth(wd)\, w^{\alpha-1}\,\mathrm{d}w = M_L + |d|^{-\alpha}[I_\alpha(L|d|) - I_\alpha(|d|/L)]$ (**??**) (substitute $s = w|d|$).*

Proof. Step 1: on '[1/L, L]', '(wd) coth(wd) w^α−1 = w^α−1 + |d|^1−α (s coth s − 1) s^α−1' with 's = |d| w'.
Step 2: the first part is 'M_L'.
Step 3: the substitution 's = |d| w'. □

Theorem 670. *For $0 < \alpha < 1$, $L > 1$, $x, x' \in [-1, 1]$ with $d = x - x' \neq 0$, $K^{(L)}(x, x') - K^{(\infty)}(x, x') = T_1 - T_2 - T_3 + T_4 + T_5$ (**??**): insert (**??**) and (**??**) into (**??**) and use $LM_L = L^{1+\alpha}(1 - L^{-2\alpha})/\alpha$.*

Proof. Step 1: the powers of 'L' and '|d|' in terms of 'A = L^α' and 'B = |d|^α'.
Step 2: clear denominators. □

Theorem 671. *For $0 < \alpha$, $L > 1$ and $x \in [-1, 1]$, $K^{(L)}(x, x) = 1 - \frac{1}{L} + T_5$ (Theorem 676; $\Psi_L(0) = M_L$).*

Proof. □

Theorem 672. *For $0 < \alpha$ and $L > 1$, $0 \leq T_3 \leq \frac{1}{L(1-L^{-2\alpha})}$ (from $J_\alpha(S) \leq S^\alpha/\alpha$).*

Proof. Step 1: '|d|^−α J_α(L|d|) ≤ |d|^−α (L|d|)^α / α = L^α / α'.
Step 2: '1 / (L (1 − L^−2α)) = (L^α / α) / (L M_L)'. □

Theorem 673. *For $0 < \alpha$ and $L > 1$, $T_3 \leq \frac{\alpha D_\alpha |d|^{-\alpha}}{L^{1+\alpha}(1-L^{-2\alpha})}$ (from $J_\alpha \leq D_\alpha$).*

Proof. □

Theorem 674. *For $0 < \alpha$, $L > 1$ and $|d| \leq 2$, $0 \leq T_4 \leq \frac{4\alpha L^{-3-2\alpha}}{3(2+\alpha)(1-L^{-2\alpha})}$ (from $I_\alpha(S) \leq \frac{S^{2+\alpha}}{3(2+\alpha)}$); this is sharper than the bound $\frac{4\alpha L^{-2-2\alpha}}{3(2+\alpha)(1-L^{-2\alpha})}$ stated in Theorem 676.*

Proof. Step 1: '|d|^−α I_α(|d|/L) ≤ |d|² L^−2−α / (3(2+α)) ≤ 4 L^−2−α / (3(2+α))'.
Step 2: divide by 'L M_L = L^1+α(1 − L^−2α)/α'. □

Theorem 675. *For $0 < \alpha < 1$, $L \geq 2$ and $x, x' \in [-1, 1]$, $0 \leq T_5 \leq \frac{2^{1-\alpha}\alpha L^{-2}}{1-L^{-2\alpha}} + \frac{2e^{-L}}{L}$ (from (**??**)).*

Proof. □

Theorem 676. *Let $K^{(\infty)}(x, x') := 1 - c_\alpha|x - x'|$. For $0 < \alpha < 1$, $L \geq 2$ and $x, x' \in [-1, 1]$, $|K^{(L)}(x, x') - K^{(\infty)}(x, x')| \leq C_\alpha L^{-\min(1,2\alpha)}$ with $C_\alpha = \frac{1}{1-2^{-2\alpha}}(2 + \frac{2\alpha}{1+\alpha} + \frac{4\alpha}{3(2+\alpha)} + 2^{1-\alpha}\alpha) + 1$ (**??**): the sum of the bounds on $T_1, \ldots, T_5$ with $1 - L^{-2\alpha} \geq 1 - 2^{-2\alpha}$, $|d| \leq 2$ and $2e^{-L} \leq 1$; on the diagonal $K^{(L)}(x, x) - 1 = -\frac{1}{L} + T_5$.*

Proof. Step 1: '1/(1 − L^−2α) ≤ Q := 1/(1 − 2^−2α)', 'Q ≥ 1'.
Step 2: 'L^−k ≤ P := L^−min(1, 2α)' for 'k ≥ min(1, 2α)'.
Step 3: the five bounds in the form '$T_i \leq c_i$ Q P'.
Step 4: sum the bounds (off the diagonal) or use 'K^(L)(x, x) = 1 − 1/L + T_5'. □

Theorem 677. *For $0 < \alpha < 1$ and $L \geq 2$, $\sup_{x,x' \in [-1,1]} |K^{(L)}(x,x') - K^{(\infty)}(x,x')| \leq C_\alpha L^{-\min(1,2\alpha)}$ (**??**).*

Proof. ☐

Theorem 678. *For $\varkappa \neq 0$ and $x \in [-1,1]$, $T[\sin \varkappa \cdot](x) = -\frac{2}{\varkappa^2} \sin \varkappa x + \frac{2\cos\varkappa}{\varkappa} x$ (proof of Theorem **??**).*

Proof. ☐

Theorem 679. *For $\varkappa \neq 0$ and $x \in [-1,1]$, $T[\cos \varkappa \cdot](x) = -\frac{2}{\varkappa^2} \cos \varkappa x + \frac{2\sin\varkappa}{\varkappa} + \frac{2\cos\varkappa}{\varkappa^2}$ (proof of Theorem **??**).*

Proof. ☐

Theorem 680. *(Theorem **??**(iii), odd case.) If $\varkappa > 0$ and $\cos \varkappa = 0$, i.e. $\varkappa \in \mathcal{K}^o = \{(j + \frac{1}{2})\pi\}$, then $K^{(\infty)}[\sin \varkappa \cdot](x) = \frac{c_\alpha}{\varkappa^2} \sin \varkappa x$ for $x \in [-1,1]$.*

Proof. ☐

Theorem 681. *(Theorem **??**(iii), even case.) If $\varkappa > 0$ and $\varkappa \sin \varkappa = \alpha \cos \varkappa$, i.e. $\varkappa \tan \varkappa = \alpha$, then $K^{(\infty)}[\cos \varkappa \cdot](x) = \frac{c_\alpha}{\varkappa^2} \cos \varkappa x$ for $x \in [-1,1]$.*

Proof. ☐

Theorem 682. *For $\alpha > 0$ and $j \geq 0$ the equation $\varkappa \tan \varkappa = \alpha$ has exactly one root $\varkappa_j^e$ in $(j\pi, j\pi + \frac{\pi}{2})$: $\varkappa \tan \varkappa$ is continuous and strictly increasing from 0 to $+\infty$ there (proof of Theorem **??**(iii)).*

Proof. ☐

Theorem 683. *$\varkappa_j^e \in (j\pi, j\pi + \frac{\alpha}{j\pi})$ for $j \geq 1$ (Theorem **??**(iii)).*

Proof. ☐

Theorem 684. *$(\varkappa_0^e)^2 \leq \alpha$ (Theorem **??**(iii)).*

Proof. ☐

Theorem 685. *(Theorem **??**(iv), (**??**).) $\frac{k\pi}{2} \leq \varkappa_{(k)} \leq \frac{(k+1)\pi}{2}$ for all $k \geq 0$, and $\varkappa_{(k)}$ is strictly increasing (interlacing $\varkappa_0^e < \frac{\pi}{2} < \varkappa_1^e < \frac{3\pi}{2} < \cdots$).*

Proof. ☐

Theorem 686. *For $k \geq 1$, $\frac{4c_\alpha}{\pi^2(k+1)^2} \leq \mu_k(K^{(\infty)}) \leq \frac{4c_\alpha}{\pi^2 k^2}$, $c_\alpha = \frac{\alpha}{1+\alpha}$ (**??**), and $\frac{1}{1+\alpha} \leq \mu_0 < 1$. In particular $\mu_k \asymp k^{-2}$, and the exponent does not depend on α.*

Proof. ☐

Theorem 687. *(Theorem **??**(i), injectivity.) If $g \in C[-1,1]$ and $K^{(\infty)} g = 0$ on $[-1,1]$, then $g = 0$: $K^{(\infty)} g = \langle g, 1 \rangle \mathbf{1} - \frac{c_\alpha}{2} T g$ with $(Tg)'' = 2g$, so differentiating twice gives $-c_\alpha g = 0$.*

Proof. Step 1: ‘(K^(∞) g)’ = 0‘ on ‘(−1, 1)‘.
Step 2: ‘(K^(∞) g)” = −c_α g = 0‘ on ‘(−1, 1)‘.
Step 3: extend to ‘[−1, 1]‘ by continuity. ☐

Theorem 688. *(Theorem **??**(ii).) If $\mu > 0$ and $g \in C[-1,1]$ satisfy $K^{(\infty)} g = \mu g$ on $[-1,1]$, then g is twice differentiable on $(-1,1)$ and $\mu\, g'' = -c_\alpha\, g$ there (**??**). (The paper states $g \in C^\infty[-1,1]$; the Lean statement records C^2 on the open interval, which is what the exhaustion uses.)*

Proof. Step 1: ‘g’ = (K^(∞) g)’/μ‘ on ‘(−1, 1)‘.

Step 2: ‘deriv g‘ coincides with ‘(K^(∞) g)’/μ‘ near every interior point, whose derivative is ‘−(c_α/μ) g‘. ☐

Theorem 689. *Let $\varkappa > 0$, $g \in C[-1,1]$ with $g' \in C[-1,1]$, $g'' = -\varkappa^2 g$ on $(-1,1)$ and $g'(\pm 1) = \mp\frac{\alpha}{2}(g(1)+g(-1))$. Then $g = A\sin\varkappa x + B\cos\varkappa x$ with $A\cos\varkappa = 0$ and $B(\varkappa\sin\varkappa - \alpha\cos\varkappa) = 0$ (proof of Theorem* **??***(iii)).*

Proof. Step 1: the difference ‘h = g − A sin ϰx − B cos ϰx‘ solves the same equation.

Step 2: the energy ‘E = h’² + ϰ² h²‘ is constant on ‘(−1, 1)‘ and vanishes at ‘0‘.

Step 3: hence ‘h = 0‘ and ‘h’ = 0‘ on ‘(−1, 1)‘, and on ‘[−1, 1]‘ by continuity.

Step 4: the boundary conditions at ‘±1‘ give the constraints on ‘A‘, ‘B‘. ☐

Theorem 690. *(Theorem* **??***(iii), exhaustion.) If $\mu > 0$ and $g \in C[-1,1]$, $g \not\equiv 0$, satisfy $K^{(\infty)}g = \mu g$ on $[-1,1]$, then $\mu = c_\alpha/\varkappa^2$ for some $\varkappa > 0$ and either $\cos\varkappa = 0$ ($\varkappa \in \mathcal{K}^o$) and $g = A\sin\varkappa x$, or $\varkappa\tan\varkappa = \alpha$ ($\varkappa \in \mathcal{K}^e$) and $g = B\cos\varkappa x$, with $A, B \neq 0$. Hence the eigenvalues and eigenfunctions are exhausted by Theorem* **??***(iii) and all eigenvalues are simple.*

Proof. Step 1: ‘g’ = (K^(∞) g)’/μ‘ on ‘(−1, 1)‘, continuous on ‘[−1, 1]‘, with ‘g” = −(c_α/μ) g = −ϰ² g‘.

Step 2: the boundary values ‘g’(±1) = ∓(α/2)(g(1) + g(−1))‘, from ‘μ(g(1) + g(−1)) = (1 − c_α) ∫ g‘ and ‘(K^(∞) g)’(±1) = ∓(c_α/2) ∫ g‘.

Step 3: ‘g = A sin ϰx + B cos ϰx‘ with ‘A cos ϰ = 0‘, ‘B(ϰ sin ϰ − α cos ϰ) = 0‘.

Step 4: ‘cos ϰ = 0‘ and ‘ϰ sin ϰ = α cos ϰ‘ cannot both hold, so exactly one of ‘A‘, ‘B‘ is nonzero. ☐

Theorem 691. *(Theorem* **??***(i), positivity.) For $g \in L^1(P_X)$, $\iint K^{(\infty)}(x,x')g(x)g(x')\,P_X(\mathrm{d}x)P_X(\mathrm{d}x') = \int\left(\int \operatorname{sgn}(x-t)g(x)\,P_X(\mathrm{d}x)\right)^2 \varpi(\mathrm{d}t) \geq 0$ (Remark 655 and Fubini).*

Proof. Step 1: ‘sgn(x − t) sgn(x’ − t) g(x) g(x’)‘ is integrable on ‘P_X ⊗ P_X ⊗ ϖ‘.

Step 2: ‘K^(∞)(x, x’) g(x) g(x’) = ∫ sgn(x − t) sgn(x’ − t) g(x) g(x’) ϖ(dt)‘ for ‘P_X ⊗ P_X‘-a.e. ‘(x, x’)‘ (‘rem:sign-network‘).

Step 3: Fubini, and the inner integral factorizes into a square. ☐

Theorem 692. *(Theorem* **??***(v).) $\sum_k \mu_k = \operatorname{Tr} K^{(\infty)} = 1$: the odd part is $\sum_j \frac{1}{(j+1/2)^2\pi^2} = \frac{1}{2}$, the even part $\sum_j (\varkappa_j^e)^{-2} = \frac{1}{\alpha} + \frac{1}{2}$ by the Hadamard factorization of $\varkappa\sin\varkappa - \alpha\cos\varkappa$.*

Proof. ☐

Theorem 693. *For bounded operators A, B on a real Hilbert space and all $k \geq 0$, $|\mu_k(A) - \mu_k(B)| \leq \|A - B\|$ for the min–max values: the min–max principle and $\langle Ag, g\rangle \leq \langle Bg, g\rangle + \|A - B\|$ (proof of Theorem* **??***(a)).*

Proof. ☐

Theorem 694. *The min–max values of $K^{(\infty)}$ on $L^2(P_X)$ are $\mu_k(K^{(\infty)}) = \frac{c_\alpha}{\varkappa_{(k)}^2}$, the eigenvalues of Theorem* **??** *in decreasing order (min–max principle for compact self-adjoint operators).*

Proof. ☐

Theorem 695. *(Theorem* **??***(a).) If $\|K^{(L)} - K^{(\infty)}\| \leq \varepsilon_L$ on $L^2(P_X)$, then $|\mu_k(K^{(L)}) - \mu_k(K^{(\infty)})| \leq \varepsilon_L$ for all $k \geq 0$.*

Proof. ☐

Theorem 696. *(Theorem **??**(b).) Let $\mu_k(K^{(L)})$ satisfy $|\mu_k(K^{(L)}) - \mu_k(K^{(\infty)})| \le \varepsilon_L$ for all k (part (a)). Then for $1 \le k \le k_{\max}(L) = \lfloor\sqrt{2c_\alpha/(\pi^2\varepsilon_L)}\rfloor - 1$, equivalently $\varepsilon_L \le \frac{2c_\alpha}{\pi^2(k+1)^2}$, $\frac{2c_\alpha}{\pi^2(k+1)^2} \le \mu_k(K^{(L)}) \le \frac{6c_\alpha}{\pi^2k^2}$* (**??**).

Proof. □

Theorem 697. *For $u \in L^1(\varpi)$ and $x \in [-\ell_\alpha, \ell_\alpha]$, $(S_\varpi u)(x) = \int \operatorname{sgn}(x-t)u(t)\,\varpi(\,\mathrm{d}t) = \frac{1}{2\ell_\alpha}\Big[\int_{-\ell_\alpha}^{x} u - \int_x^{\ell_\alpha} u\Big]$ (the integrand is u for $t < x$ and $-u$ for $t > x$).*

Proof. □

Theorem 698. *For $r \in L^1(P_X)$ and $t \in [-1,1]$, $(S^*r)(t) = \int \operatorname{sgn}(x-t)r(x)\,P_X(\,\mathrm{d}x) = \frac{1}{2}\Big[\int_t^1 r - \int_{-1}^{t} r\Big]$; for $t < -1$ it is $\frac{1}{2}\int_{-1}^{1} r$ and for $t > 1$ it is $-\frac{1}{2}\int_{-1}^{1} r$.*

Proof. □

Theorem 699. *For $u \in L^2(\varpi)$, $h := S_\varpi u$ is in $H^1(-1,1)$ with weak derivative $h' = u/\ell_\alpha$ on $(-1,1)$: $h(x) - h(-1) = \frac{1}{\ell_\alpha}\int_{-1}^{x} u$ for $x \in [-1,1]$ (Theorem **??**(ii), $\operatorname{ran} S_\varpi \subseteq H^1$).*

Proof. Both values are interval integrals over '[−ℓ_α, ℓ_α]' split at 'x' and at '−1'. □

Theorem 700. *For $f \in H^1(-1,1)$ with $f' = g$ and the representer $u_\varpi^\dagger$ of* (**??**)*, $(S_\varpi u_\varpi^\dagger)(x) = f(x)$ for every $x \in [-1,1]$: with $a = (1+\alpha)c_0$, $S_\varpi u_\varpi^\dagger(x) = \frac{1}{2\ell_\alpha}[2a(\ell_\alpha - 1) + \ell_\alpha(2f(x) - f(1) - f(-1))] = f(x)$ (Theorem **??**(ii), $H^1 \subseteq \operatorname{ran} S_\varpi$).*

Proof. Split '∫_−ℓ^x' at '−1' and '∫_x^ℓ' at '1', and evaluate the four pieces. □

Theorem 701. *$K^{(\infty)} = S_\varpi S_\varpi^*$: for $r \in L^2(P_X)$ and P_X-a.e. x, $(S_\varpi S_\varpi^* r)(x) = \int_{-1}^{1} K^{(\infty)}(x,x')r(x')\,P_X(\,\mathrm{d}x')$ with $K^{(\infty)}(x,x') = 1 - c_\alpha|x-x'|$ (Remark 655 and the integral representation of $S_\varpi S_\varpi^*$).*

Proof. □

Theorem 702. *$\operatorname{ran} S_\varpi = H^1(-1,1)$ as classes in $L^2(P_X)$: $F \in L^2(P_X)$ is of the form $S_\varpi u$ with $u \in L^2(\varpi)$ if and only if F has a representative $f \in H^1(-1,1)$. ($S_\varpi u$ is the primitive of u/ℓ_α; conversely $f = S_\varpi u_\varpi^\dagger$ with the representer* (**??**)*.)*

Proof. □

Theorem 703. *$\ker S_\varpi = \{u \in L^2(\varpi) : u = 0$ a.e. on $(-1,1)$, $\int_{-\ell_\alpha}^{-1} u = \int_1^{\ell_\alpha} u\}$. ($S_\varpi u = 0$ if and only if the H^1 function $h = S_\varpi u$ vanishes on $[-1,1]$, i.e. $h' = u/\ell_\alpha = 0$ a.e. on $(-1,1)$ and $h(1) + h(-1) = \frac{1}{\ell_\alpha}[\int_{-\ell_\alpha}^{-1} u - \int_1^{\ell_\alpha} u] = 0$; the first condition uses the Lebesgue differentiation theorem.)*

Proof. Step 1: 'S_ϖ u = 0' a.e. on '[−1, 1]', hence everywhere on '[−1, 1]' by continuity.
Step 2: the primitive '∫_−1^x u = ℓ_α (S_ϖ u (x) − S_ϖ u (−1))' vanishes on '[−1, 1]'.
Step 3: 'S_ϖ u (1) = (1/(2ℓ_α)) [∫_−ℓ^−1 u + ∫_−1^1 u − ∫_1^ℓ u] = 0'.
The condition 'u = 0' a.e. on '(−1, 1)', as a Lebesgue-a.e. statement on '[−1, 1]'.
'S_ϖ u (x) = (1/(2ℓ_α)) [∫_−ℓ^−1 u + ∫_−1^x u − ∫_x^1 u − ∫_1^ℓ u] = 0' on '[−1, 1]'. □

Theorem 704. *For $f \in H^1(-1,1)$ the minimum-norm representer is $S_\varpi^\dagger f = u_\varpi^\dagger$* (**??**)*: $u_\varpi^\dagger$ is a representer of f and is orthogonal to $\ker S_\varpi$ (by (v)), hence is the unique representer in $(\ker S_\varpi)^\perp$ (Lemma 384).*

Proof. □

Theorem 705. *For $f \in H^1(-1,1)$ with $c_0 = \frac{f(1)+f(-1)}{2}$, $\|S_\varpi^\dagger f\|^2_{L^2(\varpi)} = \|f\|^2_{\mathcal{H}_\infty} = \frac{1+\alpha}{\alpha}\|f'\|^2_{L^2(P_X)} + (1+\alpha)c_0^2$ (**??**). ($\|u_\varpi^\dagger\|^2 = \frac{1}{2\ell_\alpha}[\ell_\alpha^2 \int_{-1}^1 f'^2\,\mathrm{d}x + \frac{2\ell_\alpha^2 c_0^2}{\ell_\alpha - 1}]$ with $\ell_\alpha = \frac{1+\alpha}{\alpha}$, $\frac{\ell_\alpha}{\ell_\alpha-1} = 1+\alpha$.)*

Proof. □

Theorem 706. $\|f\|^2_{\mathcal{H}_\infty} \le \frac{1+\alpha}{\alpha}\|f'\|^2_{L^2(P_X)} + (1+\alpha)\|f\|^2_\infty$*, since* $|c_0| \le \|f\|_\infty$*.*

Proof. □

Theorem 707. $\|f\|^2_{L^2(P_X)} \le 4\|f'\|^2_{L^2(\mathrm{d}x)} + 2c_0^2$*, since* $|f(x)-c_0| \le \frac{1}{2}(|f(x)-f(1)|+|f(x)-f(-1)|) \le \frac{1}{2}\int_{-1}^1 |f'| \le \frac{1}{\sqrt{2}}\|f'\|_{L^2(\mathrm{d}x)}$*.*

Proof. Pointwise bound on '[−1, 1]': '|f(x) − c_0| ≤ A/2', hence 'f(x)^2 ≤ 2 (f x − c_0)^2 + 2c_0^2 ≤ A^2/2 + 2c_0^2 ≤ G + 2c_0^2'.

Integrate the pointwise bound against the probability measure 'P_X'. □

Theorem 708. *If $f = K^{(\infty)}g_0$ with $g_0 \in L^2(P_X)$, then $f \in H^2(-1,1)$ with $f'' = -c_\alpha g_0$, i.e. $g_0 = (K^{(\infty)})^{-1}f = -\frac{1+\alpha}{\alpha}f''$, and $f'(1) = -\frac{\alpha}{2}(f(1)+f(-1))$, $f'(-1) = +\frac{\alpha}{2}(f(1)+f(-1))$ (**??**). ($f = S_\varpi v$ with $v = S_\varpi^* g_0$, so $f' = v/\ell_\alpha$ by Theorem **??**(ii) and $v' = -g_0$; the boundary values come from $v = \pm\frac{1}{2}\int g_0$ outside $[-1,1]$.)*

Proof. 'f'' = v/ℓ_α ∈ H^1' with derivative '−g_0/ℓ_α = −c_α g_0'.

'K^(∞) g_0 = S_ϖ (S_ϖ* g_0)' and 'S_ϖ* g_0 = v' a.e.

Boundary values: 'f'(±1) = v(±1)/ℓ_α = ∓ m/(2ℓ_α)' and 'f(1) + f(−1) = (1/ℓ_α)[∫_−ℓ^−1 v − ∫_1^ℓ v] = ((ℓ_α − 1)/ℓ_α) m', 'm = ∫_−1^1 g_0'. □

Theorem 709. *If $f \in H^2(-1,1)$ with $f'(1) = -\frac{\alpha}{2}(f(1)+f(-1))$ and $f'(-1) = +\frac{\alpha}{2}(f(1)+f(-1))$, then $f = K^{(\infty)}g_0$ for $g_0 := -\frac{1+\alpha}{\alpha}f'' \in L^2(P_X)$ (**??**). ($S_\varpi^* g_0 = \ell_\alpha f' - \frac{\ell_\alpha}{2}(f'(1)+f'(-1)) = \ell_\alpha f'$ on $(-1,1)$ and $= \pm\frac{\ell_\alpha}{2}(f'(-1)-f'(1)) = \pm(1+\alpha)c_0$ outside, i.e. $S_\varpi^* g_0 = u_\varpi^\dagger$, and $S_\varpi u_\varpi^\dagger = f$ by Theorem **??**(ii).)*

Proof. 'S* G_0 = S* (−ℓ_α h)' pointwise (the analysis map only sees the 'P_X'-class).

'S* G_0 = u_ϖ†' on '(−ℓ, −1) ∪ (−1, 1) ∪ (1, ℓ)', hence 'ϖ'-a.e.

't ∈ (−ℓ_α, −1)': 'S* G_0 (t) = −(ℓ_α/2)(f'(1) − f'(−1)) = (1+α) c_0'.

't ∈ (−1, 1)': 'S* G_0 (t) = ℓ_α f'(t) − (ℓ_α/2)(f'(1) + f'(−1)) = ℓ_α f'(t)'.

't ∈ (1, ℓ_α)': 'S* G_0 (t) = (ℓ_α/2)(f'(1) − f'(−1)) = −(1+α) c_0'. □

Theorem 710. *Let (φ,ν) and (φ',ν') be two feature/hidden-law pairs on the same $L^2(P_X)$ with kernels k_ν, $k_{\nu'}$. If $|k_\nu(x,x') - k_{\nu'}(x,x')| \le \varepsilon$ for P_X-a.e. x and P_X-a.e. x', then $\|K_\nu - K_{\nu'}\|_{L^2(P_X)\to L^2(P_X)} \le \varepsilon$ (the proof of Lemma 754 only uses the bound almost everywhere).*

Proof. □

Theorem 711. *For $0<\alpha<1$ and $L \ge 2$, $\varepsilon_L := \|K^{(L)} - K^{(\infty)}\|_{L^2(P_X)\to L^2(P_X)} \le \sup_{x,x' \in [-1,1]} |K^{(L)}(x,x') - K^{(\infty)}(x,x')| \le C_\alpha L^{-\min(1,2\alpha)}$ (**??**), where $K^{(L)} = S_{\nu_0^{(L)}} S^*_{\nu_0^{(L)}}$ is the kernel operator of the tanh feature and $K^{(\infty)} = S_\varpi S_\varpi^*$ that of the sign network (Remark 655). (Theorem 676, Remark 655 and Lemma 710, since P_X-a.e. x lies in $[-1,1]$.)*

Proof. □

Theorem 712. *For $0 < \alpha < 1$, $\|K^{(L)} - K^{(\infty)}\|_{L^2(P_X)\to L^2(P_X)} \to 0$ as $L \to \infty$ ($\varepsilon_L \le C_\alpha L^{-\min(1,2\alpha)} \to 0$, Theorem 711).*

Proof. □

Theorem 713. *For $0 < \alpha < 1$, $\lambda > 0$ and $f \in L^2(P_X)$, $Q^{(L)}_\lambda(f) \to Q^{(\infty)}_\lambda(f)$ as $L \to \infty$, where $Q^{(L)}_\lambda(f) = \langle f, (K^{(L)} + \lambda)^{-1} f\rangle$ and $Q^{(\infty)}_\lambda(f) = \langle f, (K^{(\infty)} + \lambda)^{-1} f\rangle$ (Theorem 740 with $\varepsilon_L \to 0$, Theorem 712).*

Proof. □

Theorem 714. *If $f \in H^1(-1,1)$, then $Q^{(\infty)}_\lambda(f) \to \|f\|^2_{\mathcal{H}_\infty}$ as $\lambda \downarrow 0$ (Lemma 749 and Theorem* **??***(ii),(iii): $f \in \operatorname{ran} S_\varpi$ and $\|S^\dagger_\varpi f\|^2 = \|f\|^2_{\mathcal{H}_\infty}$).*

Proof. □

Theorem 715. *If $f \notin H^1(-1,1)$, then $Q^{(\infty)}_\lambda(f) \to +\infty$ as $\lambda \downarrow 0$ (Lemma 751 and Theorem* **??***(ii)).*

Proof. □

Theorem 716. *For $f \in H^1(-1,1)$ and $\lambda > 0$, $Q^{(\infty)}_\lambda(f) \le \|S^\dagger_\varpi f\|^2_{L^2(\varpi)} = \|f\|^2_{\mathcal{H}_\infty}$ (Lemma 745 and Theorem 705).*

Proof. □

Definition 717. For $\lambda > 0$ and $f \in L^2(P_X)$, $\lim_{L\to\infty} Q^{(L)}_\lambda(f)$ denotes the limit of $L \mapsto Q^{(L)}_\lambda(f)$ along $L \to \infty$ (Theorem 713: it exists and equals $Q^{(\infty)}_\lambda(f)$).

Theorem 718. *For $0 < \alpha < 1$ and $f \in L^2(P_X)$,*

$$\lim_{\lambda\downarrow 0} \lim_{L\to\infty} Q^{(L)}_\lambda(f) < \infty \iff f \in H^1(-1,1),$$

in the sense that $\lambda \mapsto \lim_{L\to\infty} Q^{(L)}_\lambda(f) = Q^{(\infty)}_\lambda(f)$ (nonincreasing in λ) is bounded on $(0,\infty)$ if and only if f has an $H^1(-1,1)$ representative. This is the precise meaning of "in the limit $L \to \infty$, (R) is equivalent to $f \in H^1$". (Theorems 713, 714, 715.)

Proof. □

Theorem 719. *For $f \in H^1(-1,1)$, $\lim_{\lambda\downarrow 0} \lim_{L\to\infty} Q^{(L)}_\lambda(f) = \|f\|^2_{\mathcal{H}_\infty}$; for $f \notin H^1(-1,1)$ the iterated limit is $+\infty$.*

Proof. □

Theorem 720. *Let $0 < \alpha < 1$, $\nu_0 = \nu^{(L)}_0$ ($L \ge 2$), (λ, β) fixed and $f \in H^1(-1,1)$, and let ρ^*_L be the corresponding minimizer of the free energy. Then*

$$\limsup_{L\to\infty} \Big[L(\rho^*_L) + \beta\, \mathrm{KL}(\rho^*_L \| \mu^{(L)}_U) \Big] \le \frac{\lambda}{2} Q^{(\infty)}_\lambda(f) \le \frac{\lambda}{2} \|f\|^2_{\mathcal{H}_\infty}$$

(**??**). *(Corollary 760 with $\|K^{(L)} - K^{(\infty)}\| \le \varepsilon_L \to 0$, Theorem 712, and Lemma 716. In Lean the reference law is $\nu^{(L)}_0$ itself, so no smoothing $\tilde{\nu}^{(L)}_0$ and no correction $1/L$ of ε_L is needed.)*

Proof. □

Theorem 721. *Under the hypotheses of Corollary 720, with ν_L^* the hidden marginal of ρ_L^* and $\kappa = \lambda/\beta$, $\limsup_L \mathrm{KL}(\nu_L^* \| \nu_0^{(L)}) \le \frac{\kappa}{2}\|f\|^2_{\mathcal{H}_\infty}$; the right-hand side does not depend on λ, β (Corollary 761).*

Proof. □

Theorem 722. *Under the hypotheses of Corollary 720, with m_L^* the conditional mean amplitude of ρ_L^*, $\limsup_L \|m_L^*\|^2_{L^2(\nu_L^*)} \le \|f\|^2_{\mathcal{H}_\infty}$: the bounds of Theorem 397 hold asymptotically with $f \in H^1$ in place of (R) and $\|f\|_{\mathcal{H}_\infty}$ in place of $\|u_0^\dagger\|$ (Corollary 762).*

Proof. □

Theorem 723. *Let $K \ge 0$ be a bounded positive operator on a Hilbert space and $x \in \overline{\operatorname{ran} K}$. Then $\lambda(K+\lambda)^{-1}x \to 0$ as $\lambda \downarrow 0$. (For $\varepsilon > 0$ write $x = (x - Kw) + Kw$ with $\|x - Kw\| < \varepsilon$; then $\|\lambda(K+\lambda)^{-1}(x-Kw)\| \le \varepsilon$ and $\|\lambda(K+\lambda)^{-1}Kw\| \le \lambda\|w\|$, Lemma 197.)*

Proof. Step 1: approximate ‘x‘ by ‘K w‘ within ‘ε/2‘.
Step 2: ‘λ R x = λ R (x − K w) + λ R (K w)‘ with the two bounds.
Step 3: ‘‖λ R x‖ ≤ ε/2 + λ‖w‖ < ε‘. □

Theorem 724. *$K^{(\infty)}$ is injective on $L^2(P_X)$: if $K^{(\infty)}r = 0$ then $\|S_\varpi^* r\|^2 = \langle K^{(\infty)}r, r\rangle = 0$, so $v := S^*r$ vanishes ϖ-a.e.; v is continuous on $[-1,1]$ with $v' = -r$ (Lemma 698), so $v = 0$ on $[-1,1]$, $\int_{-1}^x r = 0$ for all x, and $r = 0$ a.e. by Lebesgue differentiation.*

Proof. Step 1: ‘S*_ϖ r = 0‘ in ‘L²(ϖ)‘, from ‘⟪K r, r⟫ = ‖S* r‖²‘.
Step 2: ‘v = S* r‘ vanishes Lebesgue-a.e. on ‘[−1, 1]‘.
Step 3: ‘v‘ is continuous on ‘[−1, 1]‘ (an ‘H¹‘ function), so ‘v = 0‘ on ‘[−1, 1]‘.
Step 4: ‘∫_−1^x r = v(−1) − v(x) = 0‘ for all ‘x ∈ [−1, 1]‘.
Step 5: ‘r = 0‘ a.e. on ‘(−1, 1)‘, hence in ‘L²(P_X)‘. □

Theorem 725. *$\overline{\operatorname{ran} K^{(\infty)}} = (\ker K^{(\infty)})^\perp = L^2(P_X)$, since $K^{(\infty)}$ is self-adjoint and injective (Theorem 724).*

Proof. □

Theorem 726. *If $f = K^{(\infty)}g$ with $g \in L^2(P_X)$, then $R_\lambda^{(\infty)}f \to u_\infty^\dagger = S^*g$ uniformly on Z as $\lambda \downarrow 0$. ($K^{(\infty)}$ is self-adjoint and injective, so $\overline{\operatorname{ran} K^{(\infty)}} = L^2(P_X)$ and $(K^{(\infty)}+\lambda)^{-1}f = g - \lambda(K^{(\infty)}+\lambda)^{-1}g \to g$ in L^2 by Lemma 723; the bound $\|S^*r\|_\infty \le \|r\|$ of Lemma **??**(2) gives uniform convergence.)*

Proof. □

Theorem 727. *If $g = -\frac{1+\alpha}{\alpha}f''$ P_X-a.e. (i.e. $f = K^{(\infty)}g$, (**??**)), then $u_\infty^\dagger(w,b) = -\frac{1+\alpha}{2\alpha}\int_{-1}^1 \tanh(wx - b)f''(x)\,\mathrm{d}x$ (**??**): the limiting canonical ridgelet transform is the ridgelet transform, with the activation itself as filter, of the target sharpened by $-\Delta$.*

Proof. □

Theorem 728. *If $f \in H^2(-1,1)$ satisfies the boundary conditions (**??**) and $g := -\frac{1+\alpha}{\alpha}f''$, then $f = K^{(\infty)}g$, $R_\lambda^{(\infty)}f \to u_\infty^\dagger$ uniformly on Z as $\lambda \downarrow 0$, and $u_\infty^\dagger(w,b) = -\frac{1+\alpha}{2\alpha}\int_{-1}^1 \tanh(wx-b)f''(x)\,\mathrm{d}x$ (Theorems 709, 726, 727).*

Proof. □

Theorem 729. *If $f = K^{(\infty)}g$ with $g \in L^2(P_X)$, then for $L \geq 2$, $\|S_{\nu_0^{(L)}} u_\infty^\dagger - f\|_{L^2(P_X)} \leq \varepsilon_L \|g\| \leq C_\alpha L^{-\min(1,2\alpha)}\|g\|$ (**??**), since $S_{\nu_0^{(L)}} S^* g = K^{(L)}g$ and $\|K^{(L)}g - K^{(\infty)}g\| \leq \varepsilon_L\|g\|$ (Theorem 711).*

Proof. □

Theorem 730. *For $g \in L^2(P_X)$ and every L, $\|S^*g\|^2_{L^2(\nu_0^{(L)})} = \langle K^{(L)}g, g\rangle$ (**??**).*

Proof. □

Theorem 731. *If $f = K^{(\infty)}g$, then $\|u_\infty^\dagger\|^2_{L^2(\nu_0^{(L)})} = \langle K^{(L)}g, g\rangle \to \langle f, g\rangle$ as $L \to \infty$ (**??**) ($|\langle (K^{(L)} - K^{(\infty)})g, g\rangle| \leq \varepsilon_L \|g\|^2$).*

Proof. □

Theorem 732. *If $f = K^{(\infty)}g$ with $f \in H^2(-1,1)$ satisfying (**??**) and $g = -\frac{1+\alpha}{\alpha} f''$, then $\langle f, g\rangle = \langle K^{(\infty)}g, g\rangle = \|S_\varpi^* g\|^2 = \|S_\varpi^\dagger f\|^2 = \|f\|^2_{\mathcal{H}_\infty}$ (Theorem 705; $S_\varpi^* g$ is the minimum-norm representer of $f = S_\varpi S_\varpi^* g$).*

Proof. □

Theorem 733. *For $r \in L^1(P_X)$ and $t \in \mathbb{R}$, $\lim_{w\to+\infty}(S^*r)(w, wt) = \lim_{w\to+\infty} \int \tanh(w(x-t))r(x)\, P_X(\,\mathrm{d}x) = \int \operatorname{sgn}(x-t) r(x)\, P_X(\,\mathrm{d}x) = (S_\varpi^* r)(t)$, and $\lim_{w\to-\infty}(S^*r)(w, wt) = -(S_\varpi^* r)(t)$ (dominated convergence of $\tanh(w(x-t)) \to \operatorname{sgn}(w)\operatorname{sgn}(x-t)$ for $x \neq t$, with the dominating function $|r|$).*

Proof. □

Theorem 734. *For $f = K^{(\infty)}g$ and every $t \in \mathbb{R}$, $\lim_{w\to\pm\infty} u_\infty^\dagger(w, wt) = \pm(S_\varpi^* g)(t)$; that is, $u_\infty^\dagger(w, wt) \to \operatorname{sgn}(w)\, u_\varpi^\dagger(t)$, where $S_\varpi^* g = u_\varpi^\dagger$ is the representer (**??**) of the sign network (extended by the same formula to $|t| > \ell_\alpha$; the explicit values are Theorem 735). (Lemma 733.)*

Proof. □

Theorem 735. *For $f \in H^2(-1,1)$ with the boundary conditions (**??**) and $g = -\frac{1+\alpha}{\alpha} f''$: $(S_\varpi^* g)(t) = \frac{1+\alpha}{\alpha} f'(t)$ for $|t| < 1$, and $(S_\varpi^* g)(t) = \mp(1+\alpha)\frac{f(1)+f(-1)}{2}$ for $t \gtrless \pm 1$; hence $\lim_{w\to\pm\infty} u_\infty^\dagger(w, wt) = \pm\frac{1+\alpha}{\alpha} f'(t)$ for $|t| < 1$ and $\mp\operatorname{sgn}(t)(1+\alpha)\frac{f(1)+f(-1)}{2}$ for $|t| > 1$. ($\int_{-1}^1 \operatorname{sgn}(x-t) f''\,\mathrm{d}x = f'(1) + f'(-1) - 2f'(t) = -2f'(t)$ for $|t| < 1$, using $f'(1) + f'(-1) = 0$, and $\int \operatorname{sgn}(x-t) f'' = -\operatorname{sgn}(t)(f'(1) - f'(-1)) = \operatorname{sgn}(t)\alpha(f(1) + f(-1))$ for $|t| > 1$.)*

Proof. '|t| < 1': '½[−ℓ(g(1) − g(t)) + ℓ(g(t) − g(−1))] = ℓ g(t)' since 'g(1) + g(−1) = 0'.
'$t > 1$': '−½ ∫_−1^1 (−ℓ h) = (ℓ/2)(g(1) − g(−1)) = −(1+α) c_0'.
'$t < -1$': '½ ∫_−1^1 (−ℓ h) = −(ℓ/2)(g(1) − g(−1)) = (1+α) c_0'. □

Theorem 736. *For $f \in H^2(-1,1)$ with (**??**) and $g = -\frac{1+\alpha}{\alpha} f''$: for $|t| < 1$, $\lim_{w\to\pm\infty} u_\infty^\dagger(w, wt) = \pm\frac{1+\alpha}{\alpha} f'(t)$, and for $|t| > 1$, $\lim_{w\to+\infty} u_\infty^\dagger(w, wt) = -\operatorname{sgn}(t)(1+\alpha)\frac{f(1)+f(-1)}{2}$ (Theorems 734, 735).*

Proof. □

Theorem 737. *Let A_1, A_2, R_1, R_2 be bounded operators on a Hilbert space and $\lambda \in \mathbb{R}$ with $R_1(A_1 + \lambda) = \mathrm{id}$ and $(A_2 + \lambda)R_2 = \mathrm{id}$. Then $R_1 - R_2 = R_1(A_2 - A_1)R_2$.*

Proof. □

Theorem 738. *Under the hypotheses of Lemma 737 with $\lambda > 0$ and $\|R_1\|, \|R_2\| \le 1/\lambda$, $\|R_1 - R_2\| \le \|R_1\|\,\|A_1 - A_2\|\,\|R_2\| \le \lambda^{-2}\|A_1 - A_2\|$.*

Proof. □

Theorem 739. *Under the hypotheses of Lemma 738, for every f, $|\langle f, R_1 f\rangle - \langle f, R_2 f\rangle| \le \lambda^{-2}\|A_1 - A_2\|\,\|f\|^2$.*

Proof. □

Theorem 740. *Let $K_\nu = S_\nu S_\nu^*$ and $K_{\nu'} = S_{\nu'} S_{\nu'}^*$ be the kernel operators of two (feature map, hidden law) pairs on the same $L^2(P_X)$, $\lambda > 0$ and $f \in L^2(P_X)$. Then $|Q_\lambda^\nu(f) - Q_\lambda^{\nu'}(f)| \le \lambda^{-2}\|K_\nu - K_{\nu'}\|\,\|f\|^2$. In the setting of Theorem **??**, $K_\nu = K^{(L)}$, $K_{\nu'} = K^{(\infty)}$ and $\|K^{(L)} - K^{(\infty)}\| \le \varepsilon_L$, this is $|Q_\lambda^{(L)}(f) - Q_\lambda^{(\infty)}(f)| \le \lambda^{-2}\varepsilon_L\|f\|^2$. (Resolvent identity $(A+\lambda)^{-1} - (B+\lambda)^{-1} = (A+\lambda)^{-1}(B-A)(B+\lambda)^{-1}$ and $\|(K+\lambda)^{-1}\| \le \lambda^{-1}$.)*

Proof. □

Theorem 741. *For $\lambda > 0$, $f \in L^2(P_X)$ and two hidden laws $\nu, \nu' \in \mathcal{P}(Z)$, $R_{\lambda,\nu} f = S^*(K_\nu + \lambda)^{-1} f$ and $R_{\lambda,\nu'} f = S^*(K_{\nu'} + \lambda)^{-1} f$ are bounded functions on Z with $\sup_{z\in Z} |R_{\lambda,\nu} f(z) - R_{\lambda,\nu'} f(z)| \le \lambda^{-2}\|K_\nu - K_{\nu'}\|\,\|f\|$. With $\nu = \nu_0^{(L)}$, ν' the limit law and $\|K^{(L)} - K^{(\infty)}\| \le \varepsilon_L$ this is (**??**). ($\|S^* r\|_\infty \le \|r\|_{L^2(P_X)}$, Lemma **??**(2), and Theorem 740.)*

Proof. □

Theorem 742. *If $\|K_{u_L} - K_{u'}\| o0$ along a filter in L, then for every $\lambda > 0$ and $f \in L^2(P_X)$, $Q_\lambda^{u_L}(f) o Q_\lambda^{u'}(f)$ (Theorem efthm:order-of-limits-L-a).*

Proof. □

Theorem 743. *For $\lambda > 0$ and $f \in L^2(P_X)$, with $g = (K_\nu + \lambda)^{-1} f$ and $u_\lambda = S_\nu^* g = R_{\lambda,\nu} f$, $Q_\lambda^\nu(f) = \langle K_\nu g, g\rangle + \lambda\|g\|^2 = \|u_\lambda\|^2_{L^2(\nu)} + \lambda\|g\|^2_{L^2(P_X)} \ge 0$. ($f = (K_\nu + \lambda) g$ and $\langle K_\nu g, g\rangle = \|S_\nu^* g\|^2$.)*

Proof. □

Theorem 744. *For $\lambda > 0$, $f \in L^2(P_X)$ and $u_\lambda := S_\nu^*(K_\nu + \lambda)^{-1} f = R_{\lambda,\nu} f$, one has $S_\nu u_\lambda - f = -\lambda(K_\nu + \lambda)^{-1} f$ and $\frac{1}{2}\|S_\nu u_\lambda - f\|^2 + \frac{\lambda}{2}\|u_\lambda\|^2 = \frac{\lambda}{2}\langle f, (K_\nu + \lambda)^{-1}(\lambda + K_\nu)(K_\nu + \lambda)^{-1} f\rangle = \frac{\lambda}{2} Q_\lambda^\nu(f)$; by Lemma **??**(4) this is the minimum of the Tikhonov functional $u \mapsto \frac{1}{2}\|S_\nu u - f\|^2 + \frac{\lambda}{2}\|u\|^2$ over $L^2(\nu)$.*

Proof. □

Theorem 745. *For $\lambda > 0$ and every $u \in L^2(\nu)$ with $S_\nu u = f$, $Q_\lambda^\nu(f) \le \|u\|^2_{L^2(\nu)}$; under (R), $Q_\lambda^\nu(f) \le \|S_\nu^\dagger f\|^2$. (Take u in the Tikhonov minimum, Proposition 744: $\frac{\lambda}{2} Q_\lambda^\nu(f) \le \frac{1}{2}\|S_\nu u - f\|^2 + \frac{\lambda}{2}\|u\|^2 = \frac{\lambda}{2}\|u\|^2$.)*

Proof. □

Theorem 746. *For $0 < \lambda \le \lambda'$, $Q_{\lambda'}^\nu(f) \le Q_\lambda^\nu(f)$: the map $\lambda \mapsto Q_\lambda^\nu(f)$ is nonincreasing. (Without spectral theory: $Q_\lambda^\nu(f) = \min_u[\lambda^{-1}\|S_\nu u - f\|^2 + \|u\|^2]$ by Proposition 744, and the functional is pointwise nonincreasing in λ.)*

Proof. Step 1: test the ‘λ’‘-Tikhonov minimum with ‘u_λ‘.
Step 2: ‘λ’$^{-1}$ ≤ λ^{-1}‘ and the identity for ‘Q_λ‘. □

Theorem 747. *For $0 < \lambda' \leq \lambda$, $\|u_{\lambda'} - u_\lambda\|^2_{L^2(\nu)} \leq Q^\nu_{\lambda'}(f) - Q^\nu_\lambda(f)$. (The identity $J_\lambda(u) = J_\lambda(u_\lambda) + \frac{1}{2}\|S_\nu(u-u_\lambda)\|^2 + \frac{\lambda}{2}\|u-u_\lambda\|^2$ of Lemma **??**(4) with $u = u_{\lambda'}$, $J_\lambda(u_\lambda) = \frac{\lambda}{2}Q^\nu_\lambda(f)$ and $J_\lambda(u_{\lambda'}) \leq \frac{\lambda}{\lambda'}J_{\lambda'}(u_{\lambda'}) = \frac{\lambda}{2}Q^\nu_{\lambda'}(f)$.)*

Proof. '½‖S u_λ' – f‖² ≤ (λ/(2λ'))‖S u_λ' – f‖²' since 'λ' ≤ λ'. ☐

Theorem 748. *Assume (R). Then $u_\lambda = R_{\lambda,\nu}f \to S_\nu^\dagger f$ in $L^2(\nu)$ as $\lambda \downarrow 0$. (By Lemma 453, $u_\lambda - u^\dagger = -\lambda(T_0+\lambda)^{-1}u^\dagger$; $u^\dagger \in (\ker S_\nu)^\perp = \overline{\operatorname{ran} T_0}$, so for $\varepsilon > 0$ write $u^\dagger = (u^\dagger - T_0 w) + T_0 w$ with $\|u^\dagger - T_0 w\| < \varepsilon$; then $\|\lambda(T_0+\lambda)^{-1}(u^\dagger - T_0 w)\| \leq \varepsilon$ and $\|\lambda(T_0+\lambda)^{-1}T_0 w\| \leq \lambda\|w\|$.)*

Proof. Step 1: approximate 'u†' by 'T_0 w' within 'ε/2'.
Step 2: the decomposition 'u_λ – u† = $d_1 + d_2$' with '(T_0 + λ) d_1 = –λ(u† – T_0 w)' and '(T_0 + λ) d_2 = –λ T_0 w'.
Step 3: '‖u_λ – u†‖ ≤ ε/2 + λ‖w‖ < ε'. ☐

Theorem 749. *Assume (R). Then $Q^\nu_\lambda(f) \uparrow \|S_\nu^\dagger f\|^2_{L^2(\nu)}$ as $\lambda \downarrow 0$: $\lambda \mapsto Q^\nu_\lambda(f)$ is nonincreasing, bounded by $\|S_\nu^\dagger f\|^2$, and converges to it. ($Q^\nu_\lambda(f) = \langle f, g_\lambda\rangle = \langle S_\nu u^\dagger, g_\lambda\rangle = \langle u^\dagger, u_\lambda\rangle$ and $u_\lambda \to u^\dagger$, Lemma 748.)*

Proof. ☐

Theorem 750. *If $\sup_{\lambda>0} Q^\nu_\lambda(f) < \infty$, then $f \in \operatorname{ran} S_\nu$. (Let $\ell := \sup_{\lambda>0} Q^\nu_\lambda(f) = \lim_{\lambda\downarrow 0} Q^\nu_\lambda(f)$. For $0 < \lambda' \leq \lambda \leq \lambda_0$, Lemma 747 gives $\|u_{\lambda'} - u_\lambda\|^2 \leq Q_{\lambda'} - Q_\lambda \leq \ell - Q_{\lambda_0}$, so $u_{1/n}$ is Cauchy in $L^2(\nu)$ with limit u; and $\|S_\nu u_\lambda - f\|^2 \leq \lambda Q_\lambda(f) \leq \lambda \sup Q \to 0$ gives $S_\nu u = f$. No weak compactness is needed.)*

Proof. Step 1: the sequence 'v n = u_1/(n+1)' is Cauchy.
For '0 < λ' ≤ λ ≤ λ_0', '‖u_λ' – u_λ‖² ≤ Q_λ' – Q_λ ≤ ℓ – Q_λ_0 < ε²'.
Step 2: 'S_ν v n → f', since '‖S_ν v n – f‖² ≤ λ_n Q_λ_n(f) ≤ λ_n M → 0'.
Step 3: 'S_ν u = f' by uniqueness of limits. ☐

Theorem 751. *If $f \notin \operatorname{ran} S_\nu$, then $Q^\nu_\lambda(f) \uparrow +\infty$ as $\lambda \downarrow 0$: for every M there is $\lambda_0 > 0$ with $Q^\nu_\lambda(f) \geq M$ for all $0 < \lambda \leq \lambda_0$. Together with Lemma 749, $\lim_{\lambda\downarrow 0} Q^\nu_\lambda(f) < \infty$ if and only if (R) holds for ν. (Otherwise $Q^\nu_\lambda(f)$ is bounded on $(0,\infty)$ by monotonicity, and Lemma 750 gives $f \in \operatorname{ran} S_\nu$.)*

Proof. By monotonicity, 'Q_λ(f) < M' for every 'λ > 0'. ☐

Theorem 752. *For every $\nu \in \mathcal{P}(Z)$ and $f \in L^2(P_X)$, $\lambda \mapsto Q^\nu_\lambda(f)$ is nonincreasing on $(0,\infty)$, and as $\lambda \downarrow 0$, $Q^\nu_\lambda(f) \to \|S_\nu^\dagger f\|^2_{L^2(\nu)}$ if $f \in \operatorname{ran} S_\nu$, while $Q^\nu_\lambda(f) \to +\infty$ otherwise. For the limit kernel $K^{(\infty)}$ of Theorem **??**, $\|S_\nu^\dagger f\|^2 = \|f\|^2_{\mathcal{H}_\infty}$ and $\operatorname{ran} S_\nu = H^1(-1,1)$ (Theorem **??**).*

Proof. ☐

Theorem 753. *For $g \in L^2(P_X)$ and P_X-a.e. x, $(K_\nu g)(x) = \int_{\mathcal{X}} k_\nu(x,x')g(x')\,P_X(\mathrm{d}x')$ with $k_\nu(x,x') = \int_Z \varphi_z(x)\varphi_z(x')\,\nu(\mathrm{d}z)$, $|k_\nu| \leq 1$. (Fubini in $(K_\nu g)(x) = \int_Z \varphi_z(x)\int_{\mathcal{X}}\varphi_z(x')g(x')\,P_X(\mathrm{d}x')\,\nu(\mathrm{d}z)$, the integrand being dominated by $|g(x')|$.)*

Proof. ☐

Theorem 754. *Let (φ,ν) and (φ',ν') be two feature/hidden-law pairs on the same $L^2(P_X)$ with kernels $k_\nu(x,x') = \int \varphi_z(x)\varphi_z(x')\,\nu(\,\mathrm{d}z)$ and $k_{\nu'}$. If $\sup_{x,x'}|k_\nu(x,x') - k_{\nu'}(x,x')| \le \varepsilon$, then $\|K_\nu - K_{\nu'}\|_{L^2(P_X)\to L^2(P_X)} \le \varepsilon$. (By Lemma 753, $|(K_\nu g - K_{\nu'} g)(x)| \le \varepsilon\|g\|_{L^1(P_X)} \le \varepsilon\|g\|_{L^2(P_X)}$ for a.e. x, and P_X is a probability measure.) In the setting of Theorem 676, $\|K^{(L)} - K^{(\infty)}\| \le \sup_{x,x'\in[-1,1]}|K^{(L)}(x,x') - K^{(\infty)}(x,x')| \le \varepsilon_L$.*

Proof. □

Theorem 755. *Let $u \in L^2(\nu_0)$ be arbitrary and $\tilde\rho_u := \rho_{\nu_0,u} = \nu_0 \otimes \mathcal{N}(u,\beta/\lambda)$ the competitor of (**??**). Then $\tilde\rho_u \in \mathcal{D}$, $F_{\tilde\rho_u} = S_{\nu_0}u$, $L(\tilde\rho_u) = \frac{1}{2}\|S_{\nu_0}u - f\|^2$, $\mathrm{KL}(\tilde\rho_u\|\mu_U) = \frac{\lambda}{2\beta}\|u\|^2_{L^2(\nu_0)}$ and $\mathcal{F}(\tilde\rho_u) = \frac{1}{2}\|S_{\nu_0}u - f\|^2 + \frac{\lambda}{2}\|u\|^2_{L^2(\nu_0)}$ (Lemma **??** with $\nu = \nu_0$).*

Proof. Step 1: ‘KL(ρ̃_u‖μ_U) = (κ/2)‖u‖²‘.
Step 2: ‘F_ρ̃_u = S_ν₀ u‘, so ‘L(ρ̃_u) = ½‖S_ν₀ u − f‖²‘. □

Theorem 756. *Assume (A1), (A3), (A4), (A5) and $f \in L^2(P_X)$. Let ρ^* be the minimizer of $\mathcal{F}$ (Lemma **??**). For every $u \in L^2(\nu_0)$,*

$$L(\rho^*) + \beta\ \mathrm{KL}(\rho^*\|\mu_U) \le \frac{1}{2}\|S_{\nu_0}u - f\|^2_{L^2(P_X)} + \frac{\lambda}{2}\|u\|^2_{L^2(\nu_0)}.$$

(In the proof of Theorem 397, take the competitor $\tilde\rho_u = \nu_0 \otimes \mathcal{N}(u,\beta/\lambda)$ for an arbitrary u: $L(\tilde\rho_u) = \frac{1}{2}\|S_{\nu_0}u - f\|^2$ and $\mathrm{KL}(\tilde\rho_u\|\mu_U) = \frac{\lambda}{2\beta}\|u\|^2$, Lemma 755; minimality $\mathcal{F}(\rho^*) \le \mathcal{F}(\tilde\rho_u)$.)

Proof. □

Theorem 757. *Under the hypotheses of Proposition 756, the infimum of the right-hand side of (**??**) over $u \in L^2(\nu_0)$ is attained at $u = R_{\lambda,\nu_0}f$ with value $\frac{\lambda}{2}Q_\lambda^{\nu_0}(f)$ (Proposition 744), and therefore $L(\rho^*) + \beta\ \mathrm{KL}(\rho^*\|\mu_U) \le \frac{\lambda}{2}Q_\lambda^{\nu_0}(f)$.*

Proof. □

Theorem 758. *Under (R), $Q_\lambda^{\nu_0}(f) \le \|u_0^\dagger\|^2_{L^2(\nu_0)}$ (Lemma 745), and Theorem 397(a) is recovered: $L(\rho^*) + \beta\ \mathrm{KL}(\rho^*\|\mu_U) \le \frac{\lambda}{2}Q_\lambda^{\nu_0}(f) \le \frac{\lambda}{2}\|u_0^\dagger\|^2_{L^2(\nu_0)}$.*

Proof. □

Theorem 759. *Under the hypotheses of Proposition 756, with ν^* the hidden marginal of ρ^*, m^* its conditional mean amplitude and $\kappa = \lambda/\beta$,*

$$\|r^*\|^2_{L^2(P_X)} \le \lambda Q_\lambda^{\nu_0}(f), \qquad \mathrm{KL}(\rho^*\|\mu_U) \le \frac{\kappa}{2}Q_\lambda^{\nu_0}(f), \qquad \mathrm{KL}(\nu^*\|\nu_0) \le \frac{\kappa}{2}Q_\lambda^{\nu_0}(f), \qquad \|m^*\|^2_{L^2(\nu^*)} \le Q_\lambda^{\nu_0}(f),$$

and $\|m^*\|^2_{L^2(\nu^*)} + \frac{2\beta}{\lambda}\mathrm{KL}(\nu^*\|\nu_0) + \frac{2}{\lambda}L(\rho^*) \le Q_\lambda^{\nu_0}(f)$. (The chain rule $\mathrm{KL}(\rho^*\|\mu_U) = \mathrm{KL}(\nu^*\|\nu_0) + \frac{\kappa}{2}\|m^*\|^2_{L^2(\nu^*)}$ of Theorem 396, whose proof uses only the Gaussian conditional structure of Theorem 225 and not (R), substituted into Proposition 757.)

Proof. The energy bound in the form ‘2 L(ρ*) + 2β KL(ν*‖ν₀) + λ‖m*‖² ≤ λ Q‘. □

Theorem 760. *Let $(\nu_L)_L$ be hidden laws on Z whose kernel operators converge in operator norm, $\|K_{\nu_L} - K_\infty\| \to 0$, to the kernel operator $K_\infty = K_{\varphi',\nu'}$ of another feature/hidden-law pair on the same $L^2(P_X)$; let (λ, β) be fixed, $f \in L^2(P_X)$, and ρ_L^* the minimizer of the free energy with reference hidden law ν_L. Then*

$$\limsup_{L\to\infty}\Big[L(\rho_L^*) + \beta\ \mathrm{KL}(\rho_L^*\|\mu_U^{(L)})\Big] \le \frac{\lambda}{2} Q_\lambda^{(\infty)}(f), \qquad Q_\lambda^{(\infty)}(f) := \langle f, (K_\infty + \lambda)^{-1} f\rangle.$$

In the setting of Corollary 720, $\nu_L = \tilde{\nu}_0^{(L)}$, $K_\infty = K^{(\infty)}$ and $Q_\lambda^{(\infty)}(f) \le \|f\|^2_{\mathcal{H}_\infty}$ for $f \in H^1(-1,1)$. (Proposition 757 for each L and Theorem 742.)

Proof. □

Theorem 761. *Under the hypotheses of Corollary 760, with ν_L^* the hidden marginal of ρ_L^* and $\kappa = \lambda/\beta$, $\limsup_L \mathrm{KL}(\nu_L^*\|\nu_L) \le \frac{\kappa}{2} Q_\lambda^{(\infty)}(f)$ (Proposition 759).*

Proof. □

Theorem 762. *Under the hypotheses of Corollary 760, with m_L^* the conditional mean amplitude of ρ_L^*, $\limsup_L \|m_L^*\|^2_{L^2(\nu_L^*)} \le Q_\lambda^{(\infty)}(f)$ (Proposition 759); the right-hand sides of the three bounds do not depend on L.*

Proof. □

Theorem 763. *Assume (A1), (A3), (A4), (A5) and $f \in L^2(P_X)$ (no (R)). With $q := Z_*/Z_0 = \int e^{\kappa m^{*2}/2}\,\mathrm{d}\nu_0$,*

$$\nu^* = w\,\nu_0, \quad w = q^{-1}e^{\kappa m^{*2}/2}, \qquad \mathrm{KL}(\nu^*\|\nu_0) = \frac{\kappa}{2}\|m^*\|^2_{L^2(\nu^*)} - \log q, \qquad 1 \le q$$

(??). *(Theorem 226, the logarithm of the density integrated against ν^*, and $e^{\kappa m^{*2}/2} \ge 1$.)*

Proof. The exponent identity ‘κ m*2/2 = s*2/(2λβ) = −W‘.

Step 1: the density.

Step 2: the relative entropy, by the Gibbs variational identity with ‘β = 1‘.

Step 3: ‘1 ≤ q‘. □

Theorem 764. *Without (R), $1 \le q \le \exp(\frac{\kappa}{2}\|m^*\|^2_{L^2(\nu^*)}) \le \exp(\frac{\kappa}{2}Q_\lambda^{\nu_0}(f))$: $\log q = \frac{\kappa}{2}\|m^*\|^2_{L^2(\nu^*)} - \mathrm{KL}(\nu^*\|\nu_0) \le \frac{\kappa}{2}\|m^*\|^2_{L^2(\nu^*)}$ (Lemma 763) and $\|m^*\|^2_{L^2(\nu^*)} \le Q_\lambda^{\nu_0}(f)$ (Proposition 759).*

Proof. □

Theorem 765. *Without (R), $\|\nu^* - \nu_0\|_{\mathcal{M}} = \|w-1\|_{L^1(\nu_0)} \le \sqrt{2\,\mathrm{KL}(\nu^*\|\nu_0)}$ (Pinsker's inequality, Lemma 850, in the convention $\|\mu\|_{\mathcal{M}} = |\mu|(Z) = 2\|\mu\|_{\mathrm{TV}}$).*

Proof. □

Theorem 766. *Assume (A1), (A3), (A4), (A5) and $f \in L^2(P_X)$; let ρ^* be the minimizer of $\mathcal{F}$ for (λ, β, ν_0), $r^* = F_{\rho^*} - f$, ν^* its hidden marginal, m^* its conditional mean amplitude and $\kappa = \lambda/\beta$. For $u \in L^2(\nu_0)$ put $E(u) := \frac{1}{\lambda}\|S_{\nu_0}u - f\|^2_{L^2(P_X)} + \|u\|^2_{L^2(\nu_0)}$* **(??)**. *Then for every $u \in L^2(\nu_0)$,*

$$E(u) = \frac{1}{\lambda}\|r^*\|^2 + \|m^*\|^2_{L^2(\nu^*)} + \frac{1}{\lambda}\|S_{\nu_0}u - F_{\rho^*}\|^2 + \|u - m^*\|^2_{L^2(\nu_0)} + \frac{2}{\kappa}[\mathrm{KL}(\nu^*\|\nu_0) + \mathrm{KL}(\nu_0\|\nu^*)],$$

all terms on the right-hand side being nonnegative. If $S_{\nu_0} u = f$ this is (??). (The strong-convexity identity (??) of Lemma 212 at the competitor $\tilde{\rho}_u = \nu_0 \otimes \mathcal{N}(u, \beta/\lambda)$, Lemma 755, with the chain rules $\mathrm{KL}(\rho^* \| \mu_U) = \mathrm{KL}(\nu^* \| \nu_0) + \frac{\kappa}{2} \|m^*\|^2_{L^2(\nu^*)}$ and $\mathrm{KL}(\tilde{\rho}_u \| \rho^*) = \mathrm{KL}(\nu_0 \| \nu^*) + \frac{\kappa}{2} \|u - m^*\|^2_{L^2(\nu_0)}$, Lemma 390; multiply by $2/\lambda$.)

Proof. Step 1: the strong-convexity identity at the competitor ‘ρ̃_u‘.
Step 2: ‘KL(ρ*‖μ_U)‘ by the chain rule and ‘KL(ρ̃_u‖ρ*)‘ by the pair formula.
Step 3: rearrange and multiply by ‘2/λ‘. □

Theorem 767. *For every $u \in L^2(\nu_0)$,*

$$\|m^*\|^2_{L^2(\nu^*)} + \frac{1}{\lambda} \|r^*\|^2 \le E(u), \qquad \mathrm{KL}(\nu^* \| \nu_0) + \mathrm{KL}(\nu_0 \| \nu^*) \le \frac{\kappa}{2} E(u), \qquad \|u - m^*\|^2_{L^2(\nu_0)} \le E(u) - \|m^*\|^2_{L^2(\nu^*)}$$

(drop nonnegative terms in (??)).

Proof. □

Theorem 768. $\|\nu^* - \nu_0\|_{\mathcal{M}} \le \sqrt{\kappa\, Q^{\nu_0}_\lambda(f)}$ *and* $1 \le q \le e^{\kappa Q^{\nu_0}_\lambda(f)/2}$*: Pinsker's inequality* $\|\nu^* - \nu_0\|_{\mathcal{M}} \le \sqrt{2\, \mathrm{KL}(\nu^* \| \nu_0)}$ *(Lemma 765) with* $\mathrm{KL}(\nu^* \| \nu_0) \le \frac{\kappa}{2} Q^{\nu_0}_\lambda(f)$ *(Proposition 759), and Lemma 764.*

Proof. □

Theorem 769. *For $h \in L^2(P_X)$, $S^* h = S^*_{\nu^*} h$ in $L^2(\nu^*)$ and $S_{\nu^*} m^* = F_{\rho^*}$ (Theorem 231), so $\langle m^*, S^* h \rangle_{L^2(\nu^*)} = \langle F_{\rho^*}, h \rangle = \langle f, h \rangle + \langle r^*, h \rangle$.*

Proof. □

Theorem 770. *For $h \in L^2(P_X)$, with $u = S^* h$, $|u| \le \|h\|$ (Lemma 192) and $\|u\|^2_{L^2(\nu^*)} = \int u^2 w \, \mathrm{d}\nu_0 \le \|u\|^2_{L^2(\nu_0)} + \|h\|^2 \|w - 1\|_{L^1(\nu_0)}$, where $\|u\|^2_{L^2(\nu_0)} = \langle K_{\nu_0} h, h \rangle$.*

Proof. □

Theorem 771. *Under the hypotheses of Lemma 766, for every $h \in L^2(P_X)$,*

$$\|m^* - S^* h\|^2_{L^2(\nu^*)} \le \frac{1}{\lambda} \Big\| (K_{\nu_0} + \lambda) h - f \Big\|^2_{L^2(P_X)} + \|h\|^2_{L^2(P_X)} \|\nu^* - \nu_0\|_{\mathcal{M}},$$

where $\|\nu^* - \nu_0\|_{\mathcal{M}} = |\nu^* - \nu_0|(Z) = \|w - 1\|_{L^1(\nu_0)}$. (Expand $\|m^* - u\|^2_{L^2(\nu^*)}$ for $u = S^* h$; Lemma 769 for the cross term, Lemma 770 for $\|u\|^2_{L^2(\nu^*)}$ and Lemma 767 with this u for $\|m^*\|^2_{L^2(\nu^*)}$; then $-\frac{1}{\lambda}\|r^*\|^2 - 2\langle r^*, h \rangle = -\frac{1}{\lambda}\|r^* + \lambda h\|^2 + \lambda \|h\|^2 \le \lambda \|h\|^2$ and $\frac{1}{\lambda}\|K_{\nu_0} h - f\|^2 + 2\langle K_{\nu_0} h - f, h \rangle + \lambda \|h\|^2 = \frac{1}{\lambda}\|(K_{\nu_0} + \lambda) h - f\|^2$.)

Proof. Step 1: expand ‘‖m* − u‖²‘ and collect the three estimates.
Step 2: the two completions of the square. □

Theorem 772. $\|m^* - S^* h\|^2_{L^2(\nu_0)} \le q\, \|m^* - S^* h\|^2_{L^2(\nu^*)}$*, since* $\mathrm{d}\nu_0 = w^{-1}\, \mathrm{d}\nu^*$ *with* $w^{-1} = q e^{-\kappa m^{*2}/2} \le q$*.*

Proof. □

Theorem 773. *For $h = (K_{\nu_0} + \lambda)^{-1} f$ (so that $S^* h = R_{\lambda,\nu_0} f$), the first term of* (**??**) *vanishes and*

$$\|m^* - R_{\lambda,\nu_0} f\|^2_{L^2(\nu^*)} \le \|(K_{\nu_0} + \lambda)^{-1} f\|^2 \sqrt{\kappa\, Q^{\nu_0}_\lambda(f)} \le \frac{\sqrt{\kappa}}{\lambda}\, Q^{\nu_0}_\lambda(f)^{3/2},$$

since $\|\nu^* - \nu_0\|_{\mathcal{M}} \le \sqrt{\kappa Q^{\nu_0}_\lambda(f)}$ (Lemma 768) and $\lambda\|(K_{\nu_0} + \lambda)^{-1} f\|^2 \le Q^{\nu_0}_\lambda(f)$ (Lemma 743).

Proof. □

Theorem 774. *For $h \in L^2(P_X)$ and $u = S^* h$ (so $|u| \le \|h\|$), $\|\Pi\rho^* - u\,\nu_0\|_{\mathcal{M}} \le \|m^* - u\|_{L^1(\nu^*)} + \|h\|\,\|\nu^* - \nu_0\|_{\mathcal{M}} \le \|m^* - u\|_{L^2(\nu^*)} + \|h\|\,\|\nu^* - \nu_0\|_{\mathcal{M}}$, since $\Pi\rho^* = m^*\nu^*$ (Theorem 229) and $m^*\nu^* - u\nu_0 = (m^* - u)\nu^* + u(w-1)\nu_0$.*

Proof. Step 1: ‘Πρ* = m* ν*‘ and ‘u ν* = (w u) ν_0‘.
Step 2: ‘Πρ* − u ν_0 = (m* − u) ν* + (w − 1) u ν_0‘.
Step 3: the two total variations. □

Theorem 775. *Let $\nu_0 \in \mathcal{P}(Z)$, $K_\infty = K_{\varphi',\nu'}$ a second kernel operator on $L^2(P_X)$, $\delta := \|K_{\nu_0} - K_\infty\|$, $f = K_\infty g$, $G := \|g\|$, $H := \langle f, g\rangle$ and $u^\dagger_\infty := S^* g$. Then $S_{\nu_0} u^\dagger_\infty - f = (K_{\nu_0} - K_\infty) g$, $\|u^\dagger_\infty\|^2_{L^2(\nu_0)} = \langle K_{\nu_0} g, g\rangle$ and*

$$Q^{\nu_0}_\lambda(f) \le E(u^\dagger_\infty) = \frac{\eta^2}{\lambda} + \langle K_{\nu_0} g, g\rangle \le H + \delta G^2 + \frac{\delta^2 G^2}{\lambda}, \qquad \eta := \|(K_{\nu_0} - K_\infty) g\| \le \delta G$$

(**??**) *(the minimum* (**??**) *and $\langle K_{\nu_0} g, g\rangle - H = \langle (K_{\nu_0} - K_\infty) g, g\rangle \le \eta G$).*

Proof. □

Theorem 776. *Let $\nu_0 \in \mathcal{P}(Z)$ satisfy (A4), $K_\infty = K_{\varphi',\nu'}$, $\delta := \|K_{\nu_0} - K_\infty\|_{L^2(P_X)\to L^2(P_X)}$, $f = K_\infty g$, $G := \|g\|$, $u^\dagger_\infty := S^* g$; let ρ^*, ν^*, m^*, q be the quantities of the minimizer for (λ, β, ν_0) and $Q_\lambda := Q^{\nu_0}_\lambda(f)$. Then* (Amplitude.)

$$\|m^* - u^\dagger_\infty\|^2_{L^2(\nu^*)} \le G^2\Big[\Big(\frac{\delta}{\sqrt{\lambda}} + \sqrt{\lambda}\Big)^2 + \sqrt{\kappa Q_\lambda}\Big], \qquad \|m^* - u^\dagger_\infty\|^2_{L^2(\nu_0)} \le e^{\kappa Q_\lambda/2}\, G^2\Big[\Big(\frac{\delta}{\sqrt{\lambda}} + \sqrt{\lambda}\Big)^2 + \sqrt{\kappa Q_\lambda}\Big].$$

(Lemma 771 with $h = g$: $(K_{\nu_0} + \lambda) g - f = (K_{\nu_0} - K_\infty) g + \lambda g$ has norm at most $(\delta + \lambda) G$, and $\frac{1}{\lambda}(\delta + \lambda)^2 = (\delta/\sqrt{\lambda} + \sqrt{\lambda})^2$; Lemma 768 for $\|\nu^* - \nu_0\|_{\mathcal{M}}$ and q; Lemma 772.)

Proof. ‘(K_ν$_0$ + λ) g − f = (K_ν$_0$ − K_∞) g + λ g‘, of norm at most ‘(δ + λ) G‘. □

Theorem 777. (Realization, hidden marginal.) *Under the hypotheses of Theorem 776, with $H := \langle f, g\rangle$ and $\eta := \|(K_{\nu_0} - K_\infty) g\| \le \delta G$,*

$$\|F_{\rho^*} - f\|^2 \le \eta^2 + \lambda\langle K_{\nu_0} g, g\rangle \le \delta^2 G^2 + \lambda(H + \delta G^2), \qquad \mathrm{KL}(\nu^*\|\nu_0) + \mathrm{KL}(\nu_0\|\nu^*) \le \frac{\kappa}{2} E(u^\dagger_\infty),$$

where $E(u^\dagger_\infty) = \frac{\eta^2}{\lambda} + \langle K_{\nu_0} g, g\rangle$ (Lemma 767 with $u = u^\dagger_\infty$ and Theorem 775).

Proof. □

Theorem 778. (Coefficient measure.) *Under the hypotheses of Theorem 776, $\|\Pi\rho^* - u^\dagger_\infty \nu_0\|_{\mathcal{M}} \le \|m^* - u^\dagger_\infty\|_{L^2(\nu^*)} + G\sqrt{\kappa E(u^\dagger_\infty)}$, with $E(u^\dagger_\infty) = \frac{\eta^2}{\lambda} + \langle K_{\nu_0} g, g\rangle$ (Lemma 774 with $h = g$, $|u^\dagger_\infty| \le G$, and $\|\nu^* - \nu_0\|_{\mathcal{M}} \le \sqrt{2\,\mathrm{KL}(\nu^*\|\nu_0)} \le \sqrt{\kappa E(u^\dagger_\infty)}$).*

Proof. ☐

Theorem 779. *Under the hypotheses of Theorem 776, for the choice $\lambda = \delta$ one has $E(u_\infty^\dagger) \le H + 2\delta G^2$ and*

$$\|m^* - u_\infty^\dagger\|_{L^2(\nu^*)} \le G\Big[2\sqrt{\delta} + \big(\kappa(H + 2\delta G^2)\big)^{1/4}\Big],$$

and if $\kappa(H + 2\delta G^2) \le 1$ the same bound times $e^{1/4}$ holds in $L^2(\nu_0)$ **(??)**. *(For $\lambda = \delta$, $(\delta/\sqrt{\lambda} + \sqrt{\lambda})^2 = 4\delta$, $E(u_\infty^\dagger) \le H + 2\delta G^2$, $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ and $e^{\kappa Q_\lambda/2} \le e^{1/2}$.)*

Proof. Step 1: ‘Q_λ ≤ H + 2δG²‘ for ‘λ = δ‘.
Step 2: ‘(δ/√λ + √λ)² = 4λ‘ and ‘√(κ Q_λ) ≤ √(κ(H + 2δG²))‘.
Step 3: ‘√(G²(4λ + S)) ≤ G(2√λ + √S)‘.
Step 4: ‘e^κ Q_λ/2 ≤ e^1/2‘ when ‘κ(H + 2δG²) ≤ 1‘. ☐

Theorem 780. (Joint schedule.) *Let $(\nu_L)_L$ be reference measures with $\delta_L := \|K_{\nu_L} - K_\infty\| \to 0$ (for $\nu_L = \tilde{\nu}_0^{(L)}$, $\delta_L \le (C_\alpha + 1)L^{-\min(1,2\alpha)}$), $f = K_\infty g$, $u_\infty^\dagger = S^* g$, and let (λ_L, β_L) satisfy $\lambda_L \to 0$, $\kappa_L = \lambda_L/\beta_L \to 0$ and $\delta_L^2/\lambda_L \to 0$. Then, for the minimizers ρ_L^* of the free energies for $(\lambda_L, \beta_L, \nu_L)$, $m_L^* \to u_\infty^\dagger$ in $L^2(\nu_L^*)$ and in $L^2(\nu_L)$, $\|\Pi\rho_L^* - u_\infty^\dagger \nu_L\|_{\mathcal{M}} \to 0$, $F_{\rho_L^*} \to f$ in $L^2(P_X)$ and $\mathrm{KL}(\nu_L^* \| \nu_L) + \mathrm{KL}(\nu_L \| \nu_L^*) \to 0$. (Insert the schedule into Theorems 776, 777 and 778: Q_{λ_L} and $E(u_\infty^\dagger)$ stay bounded by Theorem 775, and every right-hand side tends to zero.)*

Proof. Step 1: ‘0 ≤ Q_L ≤ E_L ≤ Bnd_L → H‘, so ‘κ_L Q_L → 0‘, ‘κ_L E_L → 0‘.
Step 2: the right-hand side of Theorem (a) tends to zero.
Step 3: the coefficient measure, the realization and the relative entropies. ☐

Theorem 781. *Let $A \ge 0$ be a positive operator on a real inner product space, $\lambda > 0$, $s \ge 0$ and $x := 4s/\lambda$. Then for every z, $4\lambda\|(A+s)z\|^2 \le (1+x)\langle (A+s)(A+\lambda)^2 z, z\rangle$. (The scalar identity $(1+x)(t+\lambda)^2 - 4\lambda(t+s) = (t-\lambda)^2 + x\,t(t+2\lambda)$ gives $(1+x)\langle (A+s)(A+\lambda)^2 z, z\rangle - 4\lambda\|(A+s)z\|^2 = \langle (A+s)v, v\rangle + x\langle (A+s)A(A+2\lambda)z, z\rangle \ge 0$ with $v = (A-\lambda)z$, all operators being polynomials in A with nonnegative coefficients or squares.)*

Proof. ‘‖(A + s)z‖²‘ and the square ‘v = (A − λ)z‘.
The remaining polynomial in ‘A‘ has nonnegative coefficients. ☐

Theorem 782. *Let $\nu \in \mathcal{P}(Z)$ and $f \in L^2(P_X)$ with $Q_s^\nu(f) \le H_f$ for all $s > 0$ (for $f = K_\nu^{1/2}\tilde{h}$ one may take $H_f = \|\tilde{h}\|^2$). Then for every $\lambda > 0$, $\|(K_\nu + \lambda)^{-1} f\|^2 \le \frac{H_f}{4\lambda}$. (With $z := (K_\nu + \lambda)^{-1}(K_\nu + s)^{-1} f$ one has $(K_\nu + \lambda)^{-1} f = (K_\nu + s)z$ and $Q_s^\nu(f) = \langle (K_\nu + s)(K_\nu + \lambda)^2 z, z\rangle$, so Lemma 781 gives $4\lambda\|(K_\nu + \lambda)^{-1} f\|^2 \le (1 + 4s/\lambda)Q_s^\nu(f) \le (1 + 4s/\lambda)H_f$ for every $s > 0$; let $s \downarrow 0$.)*

Proof. Step 1: for every ‘s > 0‘, ‘4λ‖(K + λ)⁻¹ f‖² ≤ (1 + 4s/λ) Q_s(f)‘.
Step 2: let ‘s ↓ 0‘. ☐

Theorem 783. *Let ν_0, $\delta = \|K_{\nu_0} - K_\infty\|$ be as in Theorem 776 and let $f \in L^2(P_X)$ satisfy $Q_s^{(\infty)}(f) \le H_f$ for all $s > 0$ (for the sign network and $f \in H^1(-1,1)$, $H_f = \|f\|^2_{\mathcal{H}_\infty}$ by the Sobolev characterization of the limit RKHS); put $h_\lambda := (K_\infty + \lambda)^{-1} f$. Then $\|h_\lambda\| \le \frac{\sqrt{H_f}}{2\sqrt{\lambda}}$ and*

$$Q_\lambda^{\nu_0}(f) \le H_f\Big(1 + \frac{\delta}{2\lambda}\Big)^2.$$

(Lemma 782; the resolvent identity gives $Q_\lambda^{\nu_0}(f) - Q_\lambda^{(\infty)}(f) = \langle (K_{\nu_0} + \lambda)^{-1} f, (K_\infty - K_{\nu_0})h_\lambda\rangle \le \sqrt{Q_\lambda^{\nu_0}(f)/\lambda}\,\delta\,\frac{\sqrt{H_f}}{2\sqrt{\lambda}}$, and $x \le A + c\sqrt{Ax}$ implies $\sqrt{x} \le (1+c)\sqrt{A}$.)

Proof. Step 1: the resolvent identity, ‘Q^ν₀_λ − Q^∞_λ = ⟪(K_ν₀ + λ)⁻¹ f, (K_∞ − K_ν₀) h_λ⟫‘.
Step 2: with ‘x = √Q^ν₀_λ‘, ‘a = √Hf‘, ‘c = δ/(2λ)‘: ‘x² ≤ a² + c a x‘, hence ‘x ≤ (1 + c) a‘. □

Theorem 784. (Realization and hidden marginal, unconditionally.) *Under the hypotheses of Proposition 783, for the minimizer ρ^* for (λ, β, ν_0),*

$$\|F_{\rho^*}-f\| \le \sqrt{\lambda Q^{\nu_0}_\lambda(f)} \le \sqrt{H_f}\Big(\sqrt{\lambda}+\frac{\delta}{2\sqrt{\lambda}}\Big), \qquad \mathrm{KL}(\nu^*\|\nu_0)+\mathrm{KL}(\nu_0\|\nu^*) \le \frac{\kappa}{2}Q^{\nu_0}_\lambda(f), \qquad \|m^*\|^2_{L^2(\nu^*)} \le Q^{\nu_0}_\lambda(f).$$

Hence $F_{\rho^*_L} \to f$ in $L^2(P_X)$ along $\lambda_L \to 0$, $\delta_L^2/\lambda_L \to 0$ for every β, and the symmetric relative entropy tends to zero if moreover $\kappa_L \to 0$. (Proposition 759, Lemma 767 with $u = R_{\lambda,\nu_0} f$, and (a).)

Proof. ‘λ(1 + δ/(2λ))² = (√λ + δ/(2√λ))²‘. □

Theorem 785. (Amplitude, high-temperature regime.) *Under the hypotheses of Proposition 783, with $R^{(\infty)}_\lambda f = S^* h_\lambda$,*

$$\|m^* - R^{(\infty)}_\lambda f\|^2_{L^2(\nu^*)} \le \frac{H_f}{4\lambda}\Big[\frac{\delta^2}{\lambda} + \sqrt{\kappa\, Q^{\nu_0}_\lambda(f)}\Big],$$

so that $\|m^*_L - R^{(\infty)}_{\lambda_L} f\|_{L^2(\nu^*_L)} \to 0$ whenever $\delta_L/\lambda_L \to 0$ and $\sqrt{\kappa_L}/\lambda_L \to 0$. (Lemma 771 with $h = h_\lambda$: $(K_{\nu_0} + \lambda)h_\lambda - f = (K_{\nu_0} - K_\infty)h_\lambda$ has norm at most $\delta\|h_\lambda\|$; then (a).)

Proof. □

Theorem 786. *Under the hypotheses of Proposition 784, along a sequence $(\nu_L, \lambda_L, \beta_L)$ with $\lambda_L \to 0$ and $\delta_L^2/\lambda_L \to 0$, $F_{\rho^*_L} \to f$ in $L^2(P_X)$ for every choice of β_L; if moreover $\kappa_L(1 + \delta_L/(2\lambda_L))^2 \to 0$ (e.g. $\kappa_L \to 0$ with δ_L/λ_L bounded) then $\mathrm{KL}(\nu^*_L\|\nu_L) + \mathrm{KL}(\nu_L\|\nu^*_L) \to 0$.*

Proof. ‘‖r*_L‖ ≤ √Hf (√λ_L + √(δ_L²/λ_L)/2) → 0‘. □

Theorem 787. *Under the hypotheses of Proposition 785, along a sequence with $\delta_L/\lambda_L \to 0$ and $\sqrt{\kappa_L}/\lambda_L = (\lambda_L\beta_L)^{-1/2} \to 0$, $\|m^*_L - R^{(\infty)}_{\lambda_L} f\|_{L^2(\nu^*_L)} \to 0$ ((**??**) with $Q^{\nu_L}_{\lambda_L}(f) \le H_f(1 + \delta_L/(2\lambda_L))^2$ bounded).*

Proof. The bound ‘(Hf/4)(δ/λ)² + (Hf √Hf/4)(√κ/λ)(1 + δ/(2λ))‘, which tends to zero. □

Chapter 11

Material for lean-operator-ridgelet

Spectral powers of a positive operator with an orthonormal eigenbasis on a real Hilbert space (diagonal operators in a Hilbert basis, the resolvent bounds behind Tikhonov rates, the range inclusion into the orthogonal of the kernel, and the half power versus the adjoint).

Definition 788. Let $e = (e_i)_{i\in I}$ be a Hilbert basis of the real Hilbert space H and $c\colon I \to \mathbb{R}$ a bounded sequence. The *multiplication operator* by c in the coordinates of e is the bounded operator $D_c x := \sum_i c_i\langle e_i, x\rangle e_i$, of norm $\|D_c\| \le \sup_i |c_i|$; it is the unique bounded operator with $D_c e_i = c_i e_i$. (In Lean, $D_c := 0$ when c is unbounded.)

Definition 789. Let T be a bounded operator on the real Hilbert space H that is diagonal in a Hilbert basis $e = (e_i)_{i\in I}$, $Te_i = \mu_i e_i$ with $0 \le \mu_i \le M$. For $a \ge 0$ the *spectral power* T^a is the multiplication operator by $(\mu_i^a)_i$ in the coordinates of e: $T^a x := \sum_i \mu_i^a\langle e_i, x\rangle e_i$ (Definition 788), with the convention $0^a = 0$ for $a > 0$. It is the operator $f(T)$ of the continuous functional calculus for $f(t) = t^a$ on $[0, M]$.

Theorem 790. *$T^a e_i = \mu_i^a e_i$ for every i; in particular $T^1 = T$, $T^aT^b = T^{a+b}$ for $a, b > 0$ and $(T^{1/2})^2 = T$.*

Proof. □

Theorem 791. *T^a is self-adjoint and positive semidefinite, $\|T^a\| \le M^a$, and $\langle T^a x, x\rangle = \sum_i \mu_i^a\langle e_i, x\rangle^2$.*

Proof. □

Theorem 792. *For $a > 0$, $\operatorname{ran} T^a \subseteq (\ker T)^\perp = \overline{\operatorname{ran} T}$: if $Tv = 0$ then $\mu_i\langle e_i, v\rangle = 0$ for every i, so each term of $\langle v, T^a g\rangle = \sum_i \mu_i^a\langle v, e_i\rangle\langle e_i, g\rangle$ vanishes ($0^a = 0$).*

Proof. □

Theorem 793. *Let T be diagonal in the Hilbert basis e with eigenvalues $0 \le \mu_i \le M$, and $a \ge 0$. If $Tv = \lambda v$, then $T^a v = \lambda^a v$. (Coordinatewise: $\langle e_i, T^a v\rangle = \mu_i^a\langle e_i, v\rangle$ and $\langle e_i, v\rangle = 0$ unless $\mu_i = \lambda$, since eigenvectors of the self-adjoint T for distinct eigenvalues are orthogonal.)*

Proof. □

Theorem 794. *Let T be diagonal in two Hilbert bases $e = (e_i)_{i\in I}$ and $e' = (e'_j)_{j\in J}$ of H, $Te_i = \mu_i e_i$ with $0 \le \mu_i \le M$ and $Te'_j = \mu'_j e'_j$. Then for every $a \ge 0$ the spectral powers of Definition 789 defined through e and through e' coincide: by Lemma 793 both map e'_j to ${\mu'_j}^a e'_j$, and a bounded operator is determined by its values on a Hilbert basis.*

Proof. The eigenvalues along 'e'' are 'μ'_j = ⟪T e'_j, e'_j⟫ ∈ [0, ‖T‖]'. □

Theorem 795. *For $\lambda > 0$, $a > 0$ and $t \in [0,1]$, $\frac{\lambda t^a}{t+\lambda} \le \lambda^{\min(a,1)}$. For $0 < a \le 1$ the weighted arithmetic–geometric mean inequality $\lambda^{1-a}t^a \le (1-a)\lambda + at \le \lambda + t$ gives $\frac{\lambda t^a}{t+\lambda} = \lambda^a \frac{\lambda^{1-a}t^a}{t+\lambda} \le \lambda^a$; for $a > 1$, $t^a \le t$ on $[0,1]$ gives $\frac{\lambda t^a}{t+\lambda} \le \frac{\lambda t}{t+\lambda} \le \lambda$.*

Proof. □

Theorem 796. *Let T be diagonal in e with eigenvalues $0 \le \mu_i \le 1$, $\lambda > 0$, $a > 0$ and $(T+\lambda)y = \lambda T^a g$, i.e. $y = \lambda(T+\lambda)^{-1}T^a g$. Then $\|y\| \le \lambda^{\min(a,1)}\|g\|$ and $\langle Ty, y\rangle \le \left(\lambda^{\min(a+1/2,1)}\|g\|\right)^2$. (Coefficientwise $\langle e_i, y\rangle = \frac{\lambda\mu_i^a}{\mu_i+\lambda}\langle e_i, g\rangle$ and $\mu_i\langle e_i, y\rangle^2 = \left(\frac{\lambda\mu_i^{a+1/2}}{\mu_i+\lambda}\right)^2\langle e_i, g\rangle^2$; apply Lemma 795 and Parseval.)*

Proof. □

Theorem 797. *Let $T = B^*B$ be diagonal in e with eigenvalues $\mu_i \in [0, M]$. Then $\operatorname{ran} B^* \subseteq \operatorname{ran} T^{1/2}$: for g in the target space of B, $g_0 := \sum_{\mu_i \ne 0} \mu_i^{-1/2}\langle Be_i, g\rangle e_i$ satisfies $T^{1/2}g_0 = B^*g$ and $\|g_0\| \le \|g\|$ (Bessel's inequality for the orthonormal family $(\mu_i^{-1/2}Be_i)_{\mu_i \ne 0}$; $Be_i = 0$ when $\mu_i = \|Be_i\|^2 = 0$).*

Proof. Bessel: '∑_i ∈ s c_i² ≤ ‖g‖²' for every finite 's'. □

Chapter 12

Material for FoML

General lemmas on Rademacher complexities that belong upstream in FoML (lean-rademacher, `ToFoML/`): the one-sided contraction principle for an arbitrary class (from FoML's finite-class theorem), McDiarmid's inequality in the bounded-difference form, and the one-sided symmetrization and tail bounds for *separable* classes over an arbitrary index type, which relax FoML's countability (or topological separability) requirement.

Theorem 798. *Let $F = \{F_i : i \in \iota\}$ be a nonempty class of functions on $\mathcal{X}$ with $|F_i(S_k)| \leq M$ on the sample $S_1, \ldots, S_N$, and let $\psi_x \colon \mathbb{R} \to \mathbb{R}$ be L-Lipschitz for every x ($L \geq 0$). Then, in the one-sided convention, $\widehat{\mathfrak{R}}_N(\{\psi \circ F_i\}) \leq L\,\widehat{\mathfrak{R}}_N(F)$, where $(\psi \circ F_i)(x) = \psi_x(F_i(x))$. This extends FoML's finite-class contraction principle ('empiricalRademacherComplexity_without_abs_contraction_finite') to an arbitrary index set: for every sign vector σ pick a near-maximizer i_σ of $\sum_k \sigma_k \psi_{S_k}(F_i(S_k))$; the finite subclass $\{F_{i_\sigma}\}_\sigma$ realizes the left-hand side up to ε and its right-hand side is dominated by that of F.*

Proof. Near-maximizers: for every 'σ' an index 'iσ' with '⨆ i, A σ i ≤ A σ (iσ) + ε'.
FoML's finite-class contraction on the finite subclass 'H'. □

Theorem 799. *Let G be a measurable function of N i.i.d. samples $\omega_1, \ldots, \omega_N \sim \mu$ such that replacing any single ω_k changes G by at most $c > 0$. Then for every $\varepsilon \geq 0$, $\mathbb{P}(G - \mathbb{E}G \geq \varepsilon) \leq \exp(-2\varepsilon^2/(Nc^2))$ and $\mathbb{P}(G - \mathbb{E}G \leq -\varepsilon) \leq \exp(-2\varepsilon^2/(Nc^2))$ (FoML's 'mcdiarmid_inequality_pos_iid_of_const' with $t = 1/(Nc^2)$).*

Proof. □

Definition 800. A class $F = \{f_i : i \in \iota\}$ of real functions on $\mathcal{X}$ is *separable* if there is a countable $D \subseteq \iota$ such that every f_i is the pointwise limit of a sequence $(f_{u_n})_n$ with $u_n \in D$. Countable classes are separable, and so are classes with ι a separable first countable topological space and $i \mapsto f_i(x)$ continuous for every x.

Theorem 801. *Let $F = \{f_i\}$ be a class with a pointwise dense sequence $(f_{e_n})_n$ and let $\Phi \colon \iota \to \mathbb{R}$ be such that $\Phi(u_n) \to \Phi(i)$ whenever $f_{u_n} \to f_i$ pointwise. Then $\sup_{i \in \iota} \Phi(i) = \sup_n \Phi(e_n)$ (with the convention that both sides are 0 if Φ is unbounded). In particular the empirical Rademacher complexities and the uniform deviations of a bounded separable class are those of the countable subclass $\{f_{e_n}\}$.*

Proof. □

Theorem 802. *Let $F = \{f_i : i \in \iota\}$ be a nonempty separable class of measurable functions on $\mathcal{Z}$ with $|f_i| \le b$, let $\mu \in \mathcal{P}(\mathcal{Z})$, $N \ge 1$, and let $\omega_1, \dots, \omega_N \sim \mu$ be i.i.d. Then, in the one-sided convention,*

$$\mathbb{E} \sup_{i\in\iota}(\mu(f_i) - \tfrac{1}{N}\textstyle\sum_k f_i(\omega_k)) \le 2\,\mathbb{E}\,\widehat{\mathfrak{R}}_N(F), \qquad \mathbb{E}\sup_{i\in\iota}(\tfrac{1}{N}\textstyle\sum_k f_i(\omega_k) - \mu(f_i)) \le 2\,\mathbb{E}\,\widehat{\mathfrak{R}}_N(F).$$

This is the one-sided form of FoML's symmetrization ('expectation_le_rademacher', from 'symmetrization_equation') with the countability of the class relaxed to separability: all suprema are computed on a pointwise dense sequence.

Proof. □

Theorem 803. *Under the hypotheses of Lemma 802 with $b > 0$, for every $\varepsilon \ge 0$,*

$$\mathbb{P}\Big(\sup_{i\in\iota}(\mu(f_i) - \tfrac{1}{N}\textstyle\sum_k f_i(\omega_k)) \ge 2\,\mathbb{E}\,\widehat{\mathfrak{R}}_N(F) + \varepsilon\Big) \le \exp\Big(-\tfrac{N\varepsilon^2}{2b^2}\Big),$$

and the same for $\sup_i(\frac{1}{N}\sum_k f_i(\omega_k) - \mu(f_i))$. The one-sided deviation is measurable (as a supremum over the dense sequence) and has bounded differences $2b/N$, so McDiarmid's inequality (Lemma 799) applies.

Proof. □

Theorem 804. *Let $F = \{f_i : i \in \iota\}$ be a nonempty separable class of measurable functions on $\mathcal{X}$ with $|f_i| \le b$, $b > 0$, let $X\colon \Omega \to \mathcal{X}$ be measurable, $\mu \in \mathcal{P}(\Omega)$ and $\omega_1, \dots, \omega_n \sim \mu$ i.i.d. Then for every $\varepsilon \ge 0$,*

$$\mathbb{P}\Big(\sup_{i\in\iota}\Big|\tfrac{1}{n}\textstyle\sum_k f_i(X(\omega_k)) - \mathbb{E}f_i(X)\Big| \ge 2\,\mathfrak{R}_n(F) + \varepsilon\Big) \le \exp\Big(-\tfrac{n\varepsilon^2}{2b^2}\Big).$$

This is FoML's 'uniform_deviation_tail_bound_separable_of_pos' with its topological hypotheses replaced by the separability of the class.

Proof. □

Definition 805. A class $F = \{F_i : i \in \iota\}$ of real functions on $\mathcal{X}$ *pseudo-shatters* the points $(x_1, t_1), \dots, (x_n, t_n) \in \mathcal{X} \times \mathbb{R}$ if for every $T \subseteq [n]$ there is i with $t_k \le F_i(x_k) \iff k \in T$ for all k.

Definition 806. The *pseudo-dimension* of F is at most d, $\mathrm{Pdim}(F) \le d$, if F pseudo-shatters no family of $n > d$ points.

Theorem 807. *Let $g\colon \mathbb{R} \to \mathbb{R}$ be such that for every $t \in \mathbb{R}$ the level set $\{u : t \le g(u)\}$ is of the form $\{u : s \le u\}$, or $\mathbb{R}$, or $\emptyset$ (e.g. $g = \tanh$, with $s = \operatorname{artanh} t$ for $|t| < 1$). If $\mathrm{Pdim}(F) \le d$ then $\mathrm{Pdim}(g \circ F) \le d$, where $g \circ F = \{g \circ F_i : i \in \iota\}$.*

Proof. If $(x_k, t_k)_{k\le n}$ are pseudo-shattered by $g \circ F$, no level set $\{u : t_k \le g(u)\}$ is trivial (the patterns $T = \emptyset$ and $T = [n]$ would be unrealizable), so $t_k \le g(u) \iff s_k \le u$ for some s_k, and $(x_k, s_k)_{k\le n}$ are pseudo-shattered by F; hence $n \le d$. □

Definition 808. The *affine class* on E is $\mathcal{A} := \{x \mapsto \langle w, x\rangle - b : (w, b) \in E \times \mathbb{R}\}$.

Definition 809. The tanh *ridge class* on E is $\Phi := \{x \mapsto \tanh(\langle w, x\rangle - b) : (w, b) \in E \times \mathbb{R}\}$.

Theorem 810. $\mathrm{Pdim}(\mathcal{A}) \le \dim E + 1$.

Proof. □

Theorem 811. *The pseudo-dimension of the class* $\Phi = \{x \mapsto \tanh(\langle w, x\rangle - b) : (w, b) \in E \times \mathbb{R}\}$ *of* tanh *ridge functions on a* d*-dimensional inner product space* E *is at most* $d + 1$.

Proof. Since $|\tanh| < 1$, for $t \le -1$ the level set $\{u : t \le \tanh u\}$ is $\mathbb{R}$ and for $t \ge 1$ it is empty; for $|t| < 1$ it is $\{u : \operatorname{artanh} t \le u\}$ by strict monotonicity of tanh. Lemma 807 reduces to the affine class, Lemma 810. □

Theorem 812. *Let* F *be a class with* $|F_i(S_k)| \le 1$ *on the sample* $S_1, \dots, S_N$ *and* $\operatorname{Pdim}(F) \le d$. *For every* $\varepsilon > 0$ *the* $L_2(P_N)$ *covering number of* F *on the sample satisfies* $\mathcal{N}(\varepsilon, F, L_2(P_N)) \le (N(\lfloor 2/\varepsilon\rfloor + 1) + 1)^d$*; in particular* $\mathcal{N}(\varepsilon, F, L_2(P_N)) \le (4N/\varepsilon)^d$ *for* $0 < \varepsilon \le 1$, $N \ge 1$.

Proof. Round the values to the grid $-1 + j\varepsilon$, $0 \le j \le M := \lfloor 2/\varepsilon\rfloor$. Two functions with the same level sets $A_i = \{(k, j) : -1 + j\varepsilon \le F_i(S_k)\}$ differ by less than ε in every coordinate, so one representative per realized level set is an ε-cover. A set of pairs (k, j) shattered by $\{A_i\}$ is a family of points $(S_k, -1 + j\varepsilon)$ pseudo-shattered by F, so the VC dimension of $\{A_i\}$ is at most d, and the Sauer–Shelah–Perles lemma gives $|\{A_i\}| \le \sum_{k\le d} \binom{N(M+1)}{k} \le (N(M + 1) + 1)^d$. □

Theorem 813. *Let* F, G *be classes with* $|F_i(S_k)|, |G_j(S_k)| \le 1$ *on the sample. Then* $\mathcal{N}(\varepsilon, F\cdot G, L_2(P_N)) \le \mathcal{N}(\varepsilon/2, F, L_2(P_N))\, \mathcal{N}(\varepsilon/2, G, L_2(P_N))$*, where* $F \cdot G = \{F_iG_j\}$.

Proof. $|F_iG_j - F_{i'}G_{j'}| \le |F_i - F_{i'}|\,|G_j| + |F_{i'}|\,|G_j - G_{j'}| \le |F_i - F_{i'}| + |G_j - G_{j'}|$ pointwise, hence $\|F_iG_j - F_{i'}G_{j'}\|_{L_2(P_N)} \le \|F_i - F_{i'}\|_{L_2(P_N)} + \|G_j - G_{j'}\|_{L_2(P_N)}$, and the products of the centers of $\varepsilon/2$-covers form an ε-cover. □

Theorem 814. *Let* F *be a class with* $|F_i(S_k)| \le 1$ *on the sample and suppose* $\mathcal{N}(x, F, L_2(P_N)) \le (A/x)^d$ *for* $0 < x \le 1/2$, *with* $d \ge 1$, $A > 0$. *Then for every* $0 < \alpha < 1/2$, $\widehat{\mathfrak{R}}_N(F) \le 4\alpha + \frac{6}{\sqrt{N}}\sqrt{d\log(2A/\alpha)}$ *(absolute convention).*

Proof. Dudley's entropy integral in FoML's form, $\widehat{\mathfrak{R}}_N(F) \le 4\alpha + \frac{12}{\sqrt{N}}\int_\alpha^{1/2}\sqrt{\log \mathcal{N}(x, F\cup -F, L_2(P_N))}\,\mathrm{d}x$, with $\mathcal{N}(x, F \cup -F) \le 2\mathcal{N}(x, F) \le 2(A/\alpha)^d \le (2A/\alpha)^d$ on $[\alpha, 1/2]$ and $\int_\alpha^{1/2} \mathrm{d}x \le 1/2$. □

Theorem 815. *Let* F *be a class with* $|F_i(S_k)| \le 1$ *on the sample* $S_1, \dots, S_N$ *(*$N \ge 1$*) and* $\operatorname{Pdim}(F) \le d$, $d \ge 1$. *Then* $\widehat{\mathfrak{R}}_N(F) \le 18\sqrt{d\log(N + 1)/N}$.

Proof. The covering bound $\mathcal{N}(x, F, L_2(P_N)) \le (4N/x)^d$ of Lemma 812 and Lemma 814 with $\alpha = 1/(4\sqrt{N})$: $8N/\alpha = 32N\sqrt{N} \le (N + 1)^7$, so $\widehat{\mathfrak{R}}_N(F) \le (1 + 6\sqrt{7d\log(N + 1)})/\sqrt{N}$, and $1 \le 1.25\sqrt{d\log(N + 1)}$, $6\sqrt{7} \le 15.9$. □

Theorem 816. *Let* F, G *be classes with* $|F_i(S_k)|, |G_j(S_k)| \le 1$ *on the sample (*$N \ge 1$*) and* $\operatorname{Pdim}(F), \operatorname{Pdim}(G) \le d$, $d \ge 1$. *Then* $\widehat{\mathfrak{R}}_N(F \cdot G) \le 26\sqrt{d\log(N + 1)/N}$.

Proof. Lemma 813 and Lemma 812 give $\mathcal{N}(x, F \cdot G, L_2(P_N)) \le (8N/x)^{2d}$; Lemma 814 with $\alpha = 1/(4\sqrt{N})$ and $16N/\alpha = 64N\sqrt{N} \le (N + 1)^8$ gives $\widehat{\mathfrak{R}}_N(F \cdot G) \le (1 + 6\sqrt{16d\log(N + 1)})/\sqrt{N} = (1 + 24\sqrt{d\log(N + 1)})/\sqrt{N}$. □

Theorem 817. *Let* F *be a class with* $|F_i(S_k)| \le 1$ *on the sample and suppose* $\mathcal{N}(x, F, L_2(P_N)) \le (A/x)^d$ *for* $0 < x \le 1/2$, *with* $d \ge 1$ *and* $2A \ge 1$. *Then for every* $0 < \alpha < 1/2$, $\widehat{\mathfrak{R}}_N(F) \le 4\alpha + \frac{12}{\sqrt{N}}\sqrt{d}\left(\frac{1}{2}\sqrt{\log(2A)} + \sqrt{2}\right)$ *(absolute convention).*

Proof. As in Lemma 814, but the integrand is bounded pointwise: $\log \mathcal{N}(x, F\cup -F) \le d\log(2A/x) = d(\log 2A + \log(1/x))$, $\sqrt{\log(1/x)} \le \sqrt{1/x}$ (from $\log y \le y-1$), and $\int_\alpha^{1/2} x^{-1/2}\,\mathrm{d}x = 2\sqrt{1/2} - 2\sqrt{\alpha} \le \sqrt{2}$. $\square$

Theorem 818. *Let F be a class with $|F_i(S_k)| \le 1$ on the sample ($N \ge 1$) and suppose $\mathcal{N}(x, F, L_2(P_N)) \le (A/x)^d$ for $0 < x \le 1/2$, with $d \ge 1$ and $2A \ge 1$. Then $\widehat{\mathfrak{R}}_N(F) \le (1 + 6\sqrt{\log(2A)} + 12\sqrt{2})\sqrt{d/N}$.*

Proof. Lemma 817 with $\alpha = 1/(4\sqrt{N})$ and $\sqrt{d} \ge 1$. $\square$

Definition 819. A finite subset t of a pseudometric space is *ε-separated* if $d(a, b) \ge \varepsilon$ for all distinct $a, b \in t$.

Theorem 820. *Let A be a totally bounded subset of a pseudometric space and $\varepsilon > 0$. If every ε-separated finite subset of A has at most B elements, then $\mathcal{N}(\varepsilon, A) \le B$.*

Proof. Take an ε-separated subset $t \subseteq A$ of maximal cardinality $n_0 \le B$. If some $q \in A$ had $d(q, y) \ge \varepsilon$ for all $y \in t$, then $t \cup \{q\}$ would be ε-separated with $n_0 + 1$ elements. Hence $A \subseteq \bigcup_{y \in t} B(y, \varepsilon)$ and $\mathcal{N}(\varepsilon, A) \le |t| \le B$. $\square$

Theorem 821. *Let α be a nonempty finite set, P a finite index set and $B_q \subseteq \alpha$ with $|B_q| \le (1-p)|\alpha|$ for $q \in P$. If $|P|(1-p)^n < 1$, there is $K\colon [n] \to \alpha$ such that for every $q \in P$ some K_k lies outside B_q.*

Proof. Double counting: $\sum_K |\{q \in P : K([n]) \subseteq B_q\}| = \sum_{q \in P} |B_q|^n \le |P|(1-p)^n|\alpha|^n < |\alpha|^n$, so some K has no such q. $\square$

Theorem 822. *Let $|f(S_k)|, |g(S_k)| \le 1$ for all k and $\|f - g\|_{L_2(P_N)} \ge \varepsilon$, $N \ge 1$. Then $|\{k : |f(S_k) - g(S_k)| < \varepsilon/2\}| \le (1 - \varepsilon^2/8)N$.*

Proof. $\varepsilon^2 N \le \sum_k (f(S_k) - g(S_k))^2 \le 4\,|\{k : |f-g| \ge \varepsilon/2\}| + \frac{\varepsilon^2}{4}|\{k : |f-g| < \varepsilon/2\}|$, so the good coordinates number at least $\frac{3}{16}\varepsilon^2 N \ge \varepsilon^2 N/8$. $\square$

Theorem 823. *For $1 \le d \le n$, $\sum_{k \le d} \binom{n}{k} \le (en/d)^d$.*

Proof. $\square$

Theorem 824. *Let F be a class with $|F_i(S_k)| \le 1$ on the sample $S_1, \dots, S_N$ and $\mathrm{Pdim}(F) \le d$, $d \ge 1$, and let $0 < \varepsilon \le 1$. Every ε-separated finite subset of F in $L_2(P_N)$ has at most $(8/\varepsilon)^{6d}$ elements.*

Proof. Let t be ε-separated with $m = |t| \ge 1$ (for $N = 0$ all distances vanish and $m \le 1$). Put $p = \varepsilon^2/8$, $n = \lfloor 2\log m/p \rfloor + d + 1$, $M = \lfloor 4/\varepsilon \rfloor$. By Lemma 822 each pair of distinct elements of t is $\varepsilon/2$-apart on all but at most $(1-p)N$ coordinates, and $m^2(1-p)^n \le e^{2\log m - pn} < 1$, so Lemma 821 gives $K\colon [n] \to [N]$ with every pair $\varepsilon/2$-apart at some S_{K_k}. Two such functions have distinct level sets on the grid $-1 + j\varepsilon/2$, $0 \le j \le M$, over the n points S_{K_k}; the level-set family has VC dimension at most d, so $m \le \sum_{k \le d} \binom{n(M+1)}{k} \le (en(M+1)/d)^d$ (Lemma 823). With $r = m^{1/(2d)}$, $\log m = 2d\log r \le 2dr$, hence $n \le 34dr/\varepsilon^2$, $M + 1 \le 5/\varepsilon$ and $m \le (170e/\varepsilon^3)^d r^d = (170e/\varepsilon^3)^d\sqrt{m}$; thus $m \le (170e/\varepsilon^3)^{2d} \le (8/\varepsilon)^{6d}$. $\square$

Theorem 825. *Let F be a class with $|F_i(S_k)| \le 1$ on the sample $S_1, \dots, S_N$ and $\mathrm{Pdim}(F) \le d$, $d \ge 1$. Then for every $0 < \varepsilon \le 1$, $\mathcal{N}(\varepsilon, F, L_2(P_N)) \le (8/\varepsilon)^{6d}$, with no dependence on N or on the sample (Haussler's bound; the sharp form is $e(d+1)(2e/\varepsilon)^d$ in $L_1(P_N)$).*

Proof. Lemma 820 reduces the covering number to the size of an ε-separated set, bounded by Lemma 824. $\square$

Chapter 13

Material for Mathlib

General lemmas used by the formalization that belong upstream (`ToMathlib/`): the Donsker–Varadhan inequality for the Kullback–Leibler divergence, moments of the real Gaussian, the completion of the square behind the Gibbs decomposition, the total variation distance with Pinsker's inequality, and the standard facts about log-Sobolev inequalities (`ToMathlib/LogSobolev/`): the entropy functional and its variational formulas, the Holley–Stroock perturbation, the transport by proper C^1 maps, the tensorization on a Euclidean product, and the KL form (L1) by cut-off approximation.

Theorem 826. *Let $\rho \ll \mu$ be probability measures with $\log \frac{d\rho}{d\mu} \in L^1(\rho)$, and let $g \in L^1(\rho)$ with $e^g \in L^1(\mu)$. Then $\int g\, d\rho \le \int \log \frac{d\rho}{d\mu}\, d\rho + \log \int e^g\, d\mu$.*

Proof. Step 1: the tilted measure 'μ.tilted g' is a probability measure and 'ρ ≪ μ.tilted g'.
Step 2: Gibbs' inequality '0 ≤ ∫ llr ρ (μ.tilted g) dρ' for the two probability measures.
Step 3: '∫ llr ρ (μ.tilted g) dρ = ∫ llr ρ μ dρ - ∫ g dρ + log ∫ e^g dμ'. □

Theorem 827. *Let ρ, μ be probability measures with $\mathrm{KL}(\rho\|\mu) < \infty$, and let $g \in L^1(\rho)$ with $\int e^g\, d\mu < \infty$. Then $\int g\, d\rho \le \mathrm{KL}(\rho\|\mu) + \log \int e^g\, d\mu$.*

Proof. Step 1: finiteness of the divergence gives 'ρ ≪ μ' and integrability of 'llr ρ μ'.
Step 2: 'KL(ρ ‖ μ) = ∫ llr ρ μ dρ' for probability measures, and apply the llr version. □

Theorem 828. *Let ρ, μ be probability measures with $\mathrm{KL}(\rho\|\mu) < \infty$, and let $g \ge 0$ be measurable with $\int e^g\, d\mu < \infty$ (Lebesgue integral). Then $\int g\, d\rho \le \mathrm{KL}(\rho\|\mu) + \log \int e^g\, d\mu$, where the left-hand side is the Bochner integral (equal to 0 if $g \notin L^1(\rho)$).*

Proof. Step 1: 'e^g ∈ L¹(μ)', since 'e^g' is measurable with finite Lebesgue integral.
Step 2: the Bochner integral of 'e^g' is the Lebesgue integral.
Step 3: if 'g ∈ L¹(ρ)', apply the Donsker–Varadhan inequality.
Step 4: otherwise the left-hand side is '0', and '∫ e^g dμ ≥ ∫ 1 dμ = 1' gives 'log ∫ e^g dμ ≥ 0'. □

Theorem 829. *Let ρ, μ be probability measures with $\mathrm{KL}(\rho\|\mu) < \infty$, and let g be measurable with $|g| \le C$. Then $\int g\, d\rho \le \mathrm{KL}(\rho\|\mu) + \log \int e^g\, d\mu$.*

Proof. Step 1: 'g ∈ L¹(ρ)' and 'e^g ∈ L¹(μ)' (bounded by 'C' and 'e^C'). □

Theorem 830. *Let ρ, μ be probability measures with $\mathrm{KL}(\rho\|\mu) < \infty$ and let $g \ge 0$ be measurable with $\int e^g\, \mathrm{d}\mu < \infty$. Then $\int g\, \mathrm{d}\rho \le \mathrm{KL}(\rho\|\mu) + \log \int e^g\, \mathrm{d}\mu$ in $[0, \infty]$, where the left-hand side is the Lebesgue integral of g; in particular $g \in L^1(\rho)$. The proof applies the bounded Donsker–Varadhan inequality to $g \wedge n$ and lets $n \to \infty$ by monotone convergence.*

Proof. Step 1: the truncations ‘g_n = min g n‘ are bounded by ‘n‘, so the bounded Donsker–Varadhan inequality gives ‘∫ g_n dρ ≤ KL + log ∫ e^g_n dμ ≤ KL + log ∫ e^g dμ‘.

Step 2: ‘⨆ n, min (g x) n = g x‘, so by monotone convergence ‘∫⁻ g dρ = ⨆ n, ∫⁻ g_n dρ = ⨆ n, ofReal (∫ g_n dρ)‘. □

Theorem 831. *Let ρ, μ be probability measures with* $\mathrm{KL}(\rho\|\mu) < \infty$ *and let* $g \geq 0$ *be measurable with* $\int e^g \, \mathrm{d}\mu < \infty$*. Then* $g \in L^1(\rho)$*.*

Proof. □

Theorem 832. *For* $v \geq 0$ *and* $2sv < 1$*,* $x \mapsto e^{sx^2}$ *is integrable with respect to* $\mathcal{N}(0, v)$*.*

Proof. Case ‘v = 0‘: ‘$\mathcal{N}$(0, 0)‘ is the Dirac mass at ‘0‘.

Case ‘v ≠ 0‘: unfold the density and complete the square in the exponent. □

Theorem 833. *For* $v \geq 0$ *and* $2sv < 1$*,* $\int e^{sx^2} \, \mathcal{N}(0, v)(dx) = (1 - 2sv)^{-1/2}$*.*

Proof. Case ‘v = 0‘: both sides are ‘1‘.

Case ‘v ≠ 0‘: unfold the density, complete the square, and use ‘integral_gaussian‘. □

Theorem 834. *For* $v \geq 0$*,* $\int |x| \, \mathcal{N}(0, v)(dx) = \sqrt{2v/\pi}$*.*

Proof. Step 1: ‘|x| = max x 0 + max (-x) 0‘, both parts integrable.

Step 2: by symmetry ‘x ↦ -x‘ of ‘$\mathcal{N}$(0, v)‘, both parts have the same integral ‘√(v / (2π))‘ (‘integral_max_zero_gaussianReal‘).

Step 3: ‘2 √(v / (2π)) = √(2 v / π)‘. □

Theorem 835. *For* $\mu \in \mathbb{R}$ *and* $v \geq 0$*,* $\int |x| \, \mathcal{N}(\mu, v)(dx) \leq |\mu| + \sqrt{2v/\pi}$*.*

Proof. Step 1: the centered absolute moment ‘∫ |x - μ| $\mathcal{N}$(μ, v)(dx) = ∫ |y| $\mathcal{N}$(0, v)(dy)‘.

Step 2: ‘|x| ≤ |μ| + |x - μ|‘ and integrate. □

Theorem 836. *For* $\mu \in \mathbb{R}$ *and* $v \geq 0$*,* $\int x^2 \, \mathcal{N}(\mu, v)(dx) = \mu^2 + v$*.*

Proof. ‘Var = ∫ x^2 - (∫ x)^2‘ with ‘Var = v‘ and ‘∫ x = μ‘. □

Theorem 837. *For* $\lambda > 0$*,* $\beta > 0$ *and* $s \in \mathbb{R}$*,* $\int_{\mathbb{R}} e^{-(as+\lambda a^2/2)/\beta} \, da = \sqrt{2\pi\beta/\lambda} \, e^{s^2/(2\lambda\beta)}$*.*

Proof. Step 1: complete the square and pull out the constant.

Step 2: translation invariance of Lebesgue measure and ‘integral_gaussian‘. □

Theorem 838. *For* $\lambda > 0$*,* $\beta > 0$ *and* $s \in \mathbb{R}$*, as measures on* $\mathbb{R}$*,* $e^{-(as+\lambda a^2/2)/\beta} \, da = \sqrt{2\pi\beta/\lambda} \, e^{s^2/(2\lambda\beta)} \, \mathcal{N}(-s/\lambda, \beta/\lambda)(da)$*.*

Proof. Step 1: the variance ‘β/λ‘ is positive, so ‘$\mathcal{N}$(-s/λ, β/λ)‘ has the Gaussian density.

Step 2: compare the densities pointwise. □

Theorem 839. *For* $\lambda, \beta > 0$ *and* $c \in \mathbb{R}$*, as measures on* $\mathbb{R}$*,* $e^{-ac/\beta} \, \mathcal{N}(0, \beta/\lambda)(\,\mathrm{d}a) = e^{c^2/(2\lambda\beta)} \, \mathcal{N}(-c/\lambda, \beta/\lambda)(\,\mathrm{d}a)$*.*

Proof. Step 1: the product of the densities is ‘(√(2πβ/λ))$^{-1}$ e^−(a c + λ a^2/2)/β‘.

Step 2: ‘lem:gaussian-complete-square‘ and the cancellation of the constants. □

Theorem 840. *For* $v > 0$ *and* $m, m' \in \mathbb{R}$*, as measures on* $\mathbb{R}$*,* $\mathcal{N}(m, v)(\,\mathrm{d}a) = \exp\Big(\frac{a(m-m')-(m^2-m'^2)/2}{v}\Big) \mathcal{N}(m', v)(\,\mathrm{d}a)$*.*

Proof. □

Definition 841. The total variation distance of two (finite) measures μ, ν is $\|\mu - \nu\|_{\mathrm{TV}} := \sup_A |\mu(A) - \nu(A)|$, the supremum over measurable sets A.

Theorem 842. *Let μ, ν be finite measures and g a measurable function with $|g| \le C$. Then $|\int g\, d\mu - \int g\, d\nu| \le 2C\,\|\mu - \nu\|_{\mathrm{TV}}$.*

Proof. Step 1: the Hahn decomposition ‘P‘ of ‘μ - ν‘.
Step 2: ‘∫ g dμ - ∫ g dν = ∫ g d(μ - ν)|_P - ∫ g d(ν - μ)|_P^c‘.
Step 3: each of the two integrals is bounded by ‘C ‖μ - ν‖_TV‘. □

Theorem 843. *For finite measures μ, ν and a measurable map g, $\|g_*\mu - g_*\nu\|_{\mathrm{TV}} \le \|\mu - \nu\|_{\mathrm{TV}}$, since $g_*\mu(A) - g_*\nu(A) = \mu(g^{-1}A) - \nu(g^{-1}A)$.*

Proof. □

Definition 844. For a finite signed measure μ on Z, $\|\mu\|_{\mathrm{TV}} := \frac{1}{2}|\mu|(Z)$, where $|\mu| = \mu^+ + \mu^-$ is the total variation measure. For $\mu = \rho - \rho'$ with ρ, ρ' probability measures this is $\sup_A |\rho(A) - \rho'(A)|$.

Theorem 845. *If $\mu(A) - \mu(A^c) \le 2C$ for every measurable A then $\|\mu\|_{\mathrm{TV}} \le C$: on a Hahn set S of μ, $|\mu|(Z) = \mu(S^c) - \mu(S)$.*

Proof. □

Theorem 846. *For a finite measure μ and $\varphi \in L^1(\mu)$ measurable, the total variation of the signed measure $\varphi\,\mu$ is $\int |\varphi|\, \mathrm{d}\mu$ (its Jordan decomposition is $(\varphi^+\mu, \varphi^-\mu)$).*

Proof. □

Theorem 847. *With $w_i = \mathrm{d}\nu_i / \mathrm{d}(\nu_1 + \nu_2)$, $\int |w_1 - w_2|\, \mathrm{d}(\nu_1 + \nu_2) \le 2\|\nu_1 - \nu_2\|_{\mathrm{TV}}$ (split on $\{w_1 \ge w_2\}$).*

Proof. Step 1: ‘∫ |w$_1$ − w$_2$| = ∫_A (w$_1$ − w$_2$) + ∫_A^c (w$_2$ − w$_1$)‘.
Step 2: each piece is a difference of masses, bounded by the total variation distance. □

Theorem 848. *For $p \in [0,1]$ and $q \in (0,1)$, $p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \ge 2(p-q)^2$.*

Proof. Step 1: the difference ‘h p‘ of the two sides, in a form continuous on ‘[0, 1]‘.
Step 2: ‘h‘ is continuous on ‘[0, 1]‘, differentiable on ‘(0, 1)‘ with derivative ‘h’(p) = log p - log q - log (1 - p) + log (1 - q) - 4 (p - q)‘, which is monotone on ‘(0, 1)‘ and vanishes at ‘q‘.
Step 3: ‘h‘ is antitone on ‘[0, q]‘ and monotone on ‘[q, 1]‘, hence ‘h p ≥ h q = 0‘. □

Theorem 849. *Let ρ, μ be probability measures with $\mathrm{KL}(\rho\|\mu) < \infty$. Then for every measurable set A, $|\rho(A) - \mu(A)| \le \sqrt{\mathrm{KL}(\rho\|\mu)/2}$.*

Proof. Step 1: ‘ρ ≪ μ‘, ‘llr ρ μ‘ is ‘ρ‘-integrable and ‘KL(ρ ‖ μ) = ∫ klFun (dρ/dμ) dμ‘.
Step 2: Jensen’s inequality on ‘A‘ and on ‘A^c‘ (data processing along the indicator of ‘A‘): ‘∫ klFun (dρ/dμ) dμ ≥ q klFun (p/q) + (1 - q) klFun ((1 - p)/(1 - q))‘.
Step 3: the edge cases ‘q = 0‘ (then ‘p = 0‘) and ‘q = 1‘ (then ‘p = 1‘) are trivial; otherwise the binary Pinsker inequality applies. □

Theorem 850. *Let ρ, μ be probability measures with $\mathrm{KL}(\rho\|\mu) < \infty$. Then $\|\rho - \mu\|_{\mathrm{TV}} \le \sqrt{\mathrm{KL}(\rho\|\mu)/2}$.*

Proof. □

Theorem 851. *Let ρ, μ be probability measures with $\mathrm{KL}(\rho\|\mu) < \infty$ and g a measurable function with $|g| \le C$. Then $|\int g\, d\rho - \int g\, d\mu| \le 2C\sqrt{\mathrm{KL}(\rho\|\mu)/2}$.*

Proof. □

Theorem 852. *For probability measures ρ, ρ' with $\mathrm{KL}(\rho\|\rho')$ and $\mathrm{KL}(\rho'\|\rho)$ finite, $\|\rho - \rho'\|_{\mathrm{TV}}^2 \leq \frac{1}{2}\min\{\mathrm{KL}(\rho\|\rho'), \mathrm{KL}(\rho'\|\rho)\}$, hence $2\|\rho - \rho'\|_{\mathrm{TV}} \leq \sqrt{\mathrm{KL}(\rho\|\rho') + \mathrm{KL}(\rho'\|\rho)}$.*

Proof. □

Definition 853. For a finite signed measure $\gamma = \gamma^+ - \gamma^-$ (Jordan decomposition) and f integrable for $|\gamma| = \gamma^+ + \gamma^-$, $\int f \, \mathrm{d}\gamma := \int f \, \mathrm{d}\gamma^+ - \int f \, \mathrm{d}\gamma^-$.

Definition 854. The total mass of a finite signed measure is $\|\gamma\|_{\mathcal{M}} := |\gamma|(Z) = \gamma^+(Z) + \gamma^-(Z)$; it equals $2\|\gamma\|_{TV}$ for the convention $\|\gamma\|_{TV} = \sup_A |\gamma(A)|$ when $\gamma(Z) = 0$.

Theorem 855. *If $\gamma = \mu_1 - \mu_2$ with μ_1, μ_2 finite measures and f is integrable for μ_1 and μ_2, then $\int f \, \mathrm{d}\gamma = \int f \, \mathrm{d}\mu_1 - \int f \, \mathrm{d}\mu_2$.*

Proof. Step 1: ‘γ⁺ + μ₂ = γ⁻ + μ₁‘ as (finite) measures.
  Step 2: integrate ‘f‘ against both sides and rearrange. □

Theorem 856. *For $h \in L^1(\nu)$ and g measurable with $gh \in L^1(\nu)$, $\int g \, \mathrm{d}(h\nu) = \int gh \, \mathrm{d}\nu$.*

Proof. □

Theorem 857. *On a metrizable space, two finite signed measures with the same integrals $\int g \, \mathrm{d}\gamma = \int g \, \mathrm{d}\gamma'$ for every $g \in C_b$ are equal.*

Proof. Step 1: the finite measures ‘γ⁺ + γ’⁻‘ and ‘γ’⁺ + γ⁻‘ have the same integrals of bounded continuous functions, hence are equal.
  Step 2: evaluate on a measurable set. □

Theorem 858. *Let ρ, μ be probability measures with $\mathrm{KL}(\rho\|\mu) < \infty$, let A be measurable and $t \geq 0$. Then $t\,\rho(A) \leq \mathrm{KL}(\rho\|\mu) + \log(1 + (e^t - 1)\mu(A))$. (Donsker–Varadhan with $g = t\,\mathbf{1}_A$, for which $\int e^g \, \mathrm{d}\mu = 1 + (e^t - 1)\mu(A)$.) With $t = \log(1 + 1/\mu(A))$ this is $\rho(A) \leq [\mathrm{KL}(\rho\|\mu) + \log 2]/\log(1 + 1/\mu(A))$.*

Proof. Step 1: the bounded Donsker–Varadhan inequality for ‘g = t 1_A‘, ‘|g| ≤ t‘.
  Step 2: ‘∫ g dρ = t ρ(A)‘ and ‘e^g = 1 + (e^t − 1) 1_A‘, so ‘∫ e^g dμ = 1 + (e^t − 1) μ(A)‘. □

Theorem 859. *Let μ be a probability measure on a Hausdorff space such that $\{\mu\}$ is tight (e.g. any probability measure on a Polish space), and let $C \in \mathbb{R}$. Then the sublevel set $\{\rho \in \mathcal{P} : \mathrm{KL}(\rho\|\mu) \leq C\}$ is tight. Indeed, given $\varepsilon > 0$ let $t := (\max(C, 0) + \log 2)/\varepsilon$ and let K be compact with $\mu(K^c) \leq e^{-t}$; by Lemma 858, $t\rho(K^c) \leq C + \log(1 + (e^t - 1)e^{-t}) \leq C + \log 2$, so $\rho(K^c) \leq \varepsilon$.*

Proof. Step 1: the case ‘ε = ∞‘ is trivial; otherwise ‘ε’ = ε.toReal > 0‘.
  Step 2: ‘t = (max C 0 + log 2)/ε’‘ and a compact ‘K‘ with ‘μ(Kᶜ) ≤ e^−t‘.
  Step 3: for ‘ρ‘ in the sublevel set, ‘t ρ(Kᶜ) ≤ C + log 2‘, hence ‘ρ(Kᶜ) ≤ ε’‘. □

Theorem 860. *Let μ be a probability measure on a Polish space and $C \in \mathbb{R}$. Then $\{\rho \in \mathcal{P} : \mathrm{KL}(\rho\|\mu) \leq C\}$ is tight (Lemma 859, since $\{\mu\}$ is tight on a Polish space).*

Proof. □

Definition 861. For measures ρ, μ and $g \colon \alpha \to \mathbb{R}$ the *Donsker–Varadhan functional* is $\mathrm{DV}_{\rho,\mu}(g) := \int g \, \mathrm{d}\rho - \log \int e^g \, \mathrm{d}\mu$.

Theorem 862. *For probability measures ρ, μ and bounded measurable g, $\mathrm{DV}_{\rho,\mu}(g) \le \mathrm{KL}(\rho\|\mu)$ in $[0,\infty]$ (Lemma 829).*

Proof. □

Theorem 863. *For probability measures ρ, μ on a topological space with its Borel σ-algebra and $g \in C_b$, $\mathrm{DV}_{\rho,\mu}(g) \le \mathrm{KL}(\rho\|\mu)$.*

Proof. □

Theorem 864. *Let $\rho \ll \mu$ be finite measures. Then $(\log \frac{d\rho}{d\mu})^-$ is ρ-integrable: $\int (\log \frac{d\rho}{d\mu})^- \,\mathrm{d}\rho = \int r\,(\log r)^- \,\mathrm{d}\mu \le \mu(\alpha)$ with $r = \frac{d\rho}{d\mu}$, since $r(\log r)^- \le 1 - r \le 1$ on $(0,1)$.*

Proof. Step 1: ‘r (−log r)+ ≤ 1‘ for ‘r ≥ 0‘, from ‘log (1/r) ≤ 1/r − 1‘.

Step 2: transfer to ‘μ‘ through the density and bound the integrand by ‘1‘. □

Theorem 865. *Let ρ, μ be probability measures and $K \in [0,\infty]$ with $\mathrm{DV}_{\rho,\mu}(g) \le K$ for every bounded measurable g. Then $\mathrm{KL}(\rho\|\mu) \le K$. Proof: if $\rho \not\ll \mu$, the functions $t\mathbf{1}_A$ with $\mu(A) = 0 < \rho(A)$ give $\mathrm{DV} = t\rho(A) \to \infty$. Otherwise let $r = d\rho/d\mu$ and $g_n := \min(n, \log\max(r, e^{-n}))$, so that $|g_n| \le n$ and $\int e^{g_n}\,\mathrm{d}\mu \le \int (r + e^{-n})\,\mathrm{d}\mu = 1 + e^{-n}$; hence $\int g_n\,\mathrm{d}\rho \le K + \log(1+e^{-n})$. Since $\min(n, (\log r)^+) \le g_n + (\log r)^-$ and $(\log r)^- \in L^1(\rho)$ (Lemma 864), monotone convergence gives $(\log r)^+ \in L^1(\rho)$, so $\mathrm{KL}(\rho\|\mu) = \int \log r\,\mathrm{d}\rho = \lim \int g_n\,\mathrm{d}\rho \le K$ by dominated convergence.*

Proof. Step 0: the case ‘K = ∞‘ is trivial.

Step 1: ‘ρ ≪ μ‘; otherwise a ‘μ‘-null set ‘A‘ with ‘ρ(A) > 0‘ and the test functions ‘t 1_A‘, for which ‘DV = t ρ(A)‘, show that ‘K = ∞‘.

Step 2: the truncations ‘g_n = min n (log (max r e^−n))‘ of the log-likelihood ratio, with ‘$|g_n|$ ≤ n‘ and ‘e^g_n ≤ r + e^−n‘, so that ‘∫ e^g_n dμ ≤ 1 + e^−n‘.

Step 3: the bound ‘∫ g_n dρ ≤ K + log (1 + e^−n)‘.

Step 4: ‘ρ‘-a.e. ‘r > 0‘, and there ‘g_n = min n (max (log r) (−n))‘ is the clamped log-likelihood ratio: ‘$|g_n|$ ≤ |log r|‘, ‘g_n → log r‘, ‘min n (log r)+ ≤ g_n + (log r)−‘.

Step 5: the negative part of ‘log r‘ is ‘ρ‘-integrable, and the uniform bound of Step 3 gives, by monotone convergence, that the positive part is ‘ρ‘-integrable too.

Step 6: ‘KL(ρ‖μ) = ∫ log r dρ = lim ∫ g_n dρ ≤ K‘ by dominated convergence (with the bound ‘|log r|‘), since ‘∫ g_n dρ ≤ K + log (1 + e^−n) → K‘. □

Theorem 866. *Let ρ, μ be probability measures on a metrizable space with its Borel σ-algebra, and $K \in [0,\infty]$ with $\mathrm{DV}_{\rho,\mu}(g) \le K$ for every $g \in C_b$. Then $\mathrm{DV}_{\rho,\mu}(g) \le K$ for every bounded measurable g. Proof: given $|g| \le C$ and $\varepsilon > 0$, choose $g_1 \in C_b$ with $\int |g - g_1|\,\mathrm{d}(\rho+\mu) \le \varepsilon$ (density of C_b in L^1 of a finite measure on a metrizable space) and clamp it to $g_2 := \max(-C, \min(C, g_1)) \in C_b$, which still satisfies $\int |g - g_2|\,\mathrm{d}(\rho+\mu) \le \varepsilon$. Then $|\int g\,\mathrm{d}\rho - \int g_2\,\mathrm{d}\rho| \le \varepsilon$ and $|\int e^g\,\mathrm{d}\mu - \int e^{g_2}\,\mathrm{d}\mu| \le e^C\varepsilon$, so $\mathrm{DV}(g) \le \mathrm{DV}(g_2) + \varepsilon + \log(\int e^g\,\mathrm{d}\mu + e^C\varepsilon) - \log\int e^g\,\mathrm{d}\mu$, and the error tends to 0 as $\varepsilon \to 0$.*

Proof. Step 0: the case ‘K = ∞‘ is trivial; otherwise it suffices to show ‘DV(g) ≤ K + δ‘ for every ‘δ > 0‘.

Step 1: choose ‘ε > 0‘ with ‘ε + log (I + e^C ε) < log I + δ‘, ‘I = ∫ e^g dμ‘.

Step 2: a bounded continuous ‘g_1‘ with ‘∫ |g − g_1| d(ρ + μ) ≤ ε‘, clamped to ‘[−C, C]‘ as ‘g_2‘; then ‘∫ |g − g_2| dρ ≤ ε‘ and ‘∫ |g − g_2| dμ ≤ ε‘.

Step 3: ‘∫ g_2 dρ ≥ ∫ g dρ − ε‘ and ‘∫ e^g_2 dμ ≤ I + e^C ε‘.

Step 4: combine with ‘DV(g_2) ≤ K‘. □

Theorem 867. *Let ρ, μ be probability measures on a metrizable space with its Borel σ-algebra and $K \in [0, \infty]$ with $\mathrm{DV}_{\rho,\mu}(g) \leq K$ for every $g \in C_b$. Then $\mathrm{KL}(\rho\|\mu) \leq K$ (Lemmas 866 and 865).*

Proof. □

Theorem 868. *Let ρ, μ be probability measures on a metrizable space with its Borel σ-algebra. Then $\mathrm{KL}(\rho\|\mu) = \sup_{g \in C_b} \left[\int g \, \mathrm{d}\rho - \log \int e^g \, \mathrm{d}\mu\right]$ in $[0, \infty]$ (the supremum of the nonnegative parts; the inequality $\geq$ is the Donsker–Varadhan inequality, and $\leq$ is Lemma 867).*

Proof. □

Theorem 869. *Let μ be a probability measure on a metrizable space with its Borel σ-algebra, and let $\rho_i \to \rho$ weakly (along a filter) in $\mathcal{P}$. Then $\mathrm{KL}(\rho\|\mu) \leq \liminf_i \mathrm{KL}(\rho_i\|\mu)$. Proof: for $g \in C_b$, $\mathrm{DV}_{\rho,\mu}(g) = \lim_i \mathrm{DV}_{\rho_i,\mu}(g) \leq \liminf_i \mathrm{KL}(\rho_i\|\mu)$ by the Donsker–Varadhan inequality and the weak convergence, and the supremum over g is $\mathrm{KL}(\rho\|\mu)$ by Lemma 868.*

Proof. ‘DV_ρ$_i$(g) → DV_ρ(g)‘ by weak convergence, and ‘DV_ρ$_i$(g) ≤ KL(ρ$_i$‖μ)‘. □

Theorem 870. *Let μ_i be σ-finite measures and $f_i \geq 0$ be μ_i-integrable. Then $\bigotimes_i (f_i \mu_i) = \left(\prod_i f_i(x_i)\right) \bigotimes_i \mu_i$. (Both sides agree on measurable boxes by Fubini.)*

Proof. □

Theorem 871. *For a σ-finite μ and a permutation σ of the finite index set, the image of $\mu^{\otimes\iota}$ under $\theta \mapsto \theta \circ \sigma$ is $\mu^{\otimes\iota}$.*

Proof. □

Theorem 872. *If T is measurable with $T_*\nu = \nu$ and $D \circ T = D$ for a measurable $D \geq 0$, then $T_*(D\nu) = D\nu$.*

Proof. □

Theorem 873. *For a σ-finite ρ, measurable Y and $t \in \mathbb{R}$, $\int e^{t \sum_i Y(x_i)} \, \rho^{\otimes\iota}(\mathrm{d}x) = \left(\int e^{tY} \, \mathrm{d}\rho\right)^{|\iota|}$ (Fubini; no integrability is needed, both sides being 0 when the right-hand factor is not integrable).*

Proof. □

Theorem 874. *For all $x \in \mathbb{R}$, $e^x \leq 1 + x + x^2 e^{|x|}$. (For $|x| \leq 1$ this is the Taylor bound $|e^x - 1 - x| \leq x^2$; for $x \geq 1$, $e^x \leq x^2 e^x$; for $x \leq -1$, $e^x \leq 1 \leq 1 + x + x^2$.)*

Proof. □

Theorem 875. *Let $\rho \in \mathcal{P}(\Omega)$ and Y be measurable with $\mathbb{E}_\rho Y = 0$ and $K := \mathbb{E}_\rho[Y^2 e^{|Y|}] < \infty$. Then for $|u| \leq 1$, $\mathbb{E}_\rho e^{uY} \leq 1 + Ku^2$, hence $\log \mathbb{E}_\rho e^{uY} \leq Ku^2$. (Integrate $e^{uY} \leq 1 + uY + u^2 Y^2 e^{|Y|}$.)*

Proof. ‘|Y| ≤ 1 + Y² ≤ 1 + Y² e^|Y|‘, so ‘Y ∈ L¹(ρ)‘.
The pointwise bound ‘e^uY ≤ 1 + uY + u² Y² e^|Y|‘ for ‘|u| ≤ 1‘. □

Theorem 876. *For σ-finite μ, ν and a measurable isomorphism e, $\mathrm{KL}(e_*\mu\|e_*\nu) = \mathrm{KL}(\mu\|\nu)$.*

Proof. □

Theorem 877. *Let ρ be a finite measure on $\alpha \times \Omega$ with Ω standard Borel, ν_1 finite and ν_2 a probability measure. Then $\mathrm{KL}(\rho^{(1)}\|\nu_1) \leq \mathrm{KL}(\rho\|\nu_1 \otimes \nu_2)$, where $\rho^{(1)}$ is the first marginal. (Chain rule: $\mathrm{KL}(\rho\|\nu_1 \otimes \nu_2) = \mathrm{KL}(\rho^{(1)}\|\nu_1) + \mathrm{KL}(\rho\|\rho^{(1)} \otimes \nu_2)$ with $\rho = \rho^{(1)} \otimes_{\mathrm{m}} \kappa$ its disintegration.)*

Proof. □

Theorem 878. *Under the hypotheses of Lemma 877 with the roles of the factors exchanged (α standard Borel, ν_1 a probability measure), $\mathrm{KL}(\rho^{(2)}\|\nu_2) \le \mathrm{KL}(\rho\|\nu_1 \otimes \nu_2)$.*

Proof. □

Theorem 879. *Let Ω be standard Borel, $\nu \in \mathcal{P}(\Omega)$ and $\mu \in \mathcal{P}(\Omega^n)$ with marginals $\mu^{(i)}$. Then $\sum_{i=1}^n \mathrm{KL}(\mu^{(i)}\|\nu) \le \mathrm{KL}(\mu\|\nu^{\otimes n})$. (Induction on n with the chain rule and Lemmas 877, 878.)*

Proof. □

Theorem 880. *Let E be a real inner product space of dimension d and let $n \ge d+2$. For all points $(x_i, s_i) \in E \times \mathbb{R}$, $i = 1, \dots, n$, there is a pattern $T \subseteq [n]$ such that no $(w, b) \in E \times \mathbb{R}$ satisfies $s_i \le \langle w, x_i\rangle - b \iff i \in T$ for all i. In other words, the affine class $\{x \mapsto \langle w, x\rangle - b\}$ has pseudo-dimension at most $d+1$.*

Proof. The $n > d+1$ vectors $(x_i, 1) \in E \times \mathbb{R}$ are linearly dependent: there is $c \ne 0$ with $\sum_i c_i x_i = 0$ and $\sum_i c_i = 0$, hence $\sum_i c_i(\langle w, x_i\rangle - b - s_i) = -\kappa$ with $\kappa := \sum_i c_i s_i$ for all (w, b). Replacing c by $-c$ we may assume $\kappa \ge 0$. Since $c \ne 0$ sums to zero, some c_{i_0} is negative. If (w, b) realized the pattern $T = \{i : c_i \ge 0\}$, every term $c_i(\langle w, x_i\rangle - b - s_i)$ would be nonnegative and the term i_0 positive, so the sum would be positive, contradicting $-\kappa \le 0$. □

Theorem 881. *Let E be a real inner product space of dimension d and let $S \subset E$ have $d+2$ points. There is $T \subseteq S$ such that for no $(w, c) \in E \times \mathbb{R}$ one has $\{v \in S : \langle w, v\rangle + c > 0\} = T$. Hence the VC dimension of the class of affine halfspaces of E is at most $d+1$.*

Proof. Enumerate $S = \{x_1, \dots, x_{d+2}\}$. A halfspace pattern $\langle w, x_i\rangle + c > 0 \iff i \in T$ is the complementary affine pattern $0 \le \langle -w, x_i\rangle - c \iff i \notin T$, so shattering S by halfspaces would pseudo-shatter the points $(x_i, 0)$ by the affine class, contradicting Lemma 880. □

Definition 882. For a measure μ and $f \ge 0$ with $f,\ f\log f \in L^1(\mu)$,

$$\mathrm{Ent}_\mu(f) := \int f \log f \,\mathrm{d}\mu - \Big(\int f \,\mathrm{d}\mu\Big) \log\Big(\int f \,\mathrm{d}\mu\Big)$$

(with $0 \log 0 = 0$).

Theorem 883. *μ satisfies $\mathrm{LSI}(\alpha)$ if and only if $\mathrm{Ent}_\mu(g^2) \le \frac{2}{\alpha} \int |\nabla g|^2 \,\mathrm{d}\mu$ for all $g \in C_c^1$.*

Proof. □

Theorem 884. *For $u \ge 0$ and $v \in \mathbb{R}$, $uv \le u \log u - u + e^v$.*

Proof. □

Theorem 885. *For $u \ge 0$ and $t > 0$, $u \log u - u \log t - u + t \ge 0$.*

Proof. □

Theorem 886. *For $0 \le u \le v$, $|u \log u| \le 1 + |v \log v|$.*

Proof. □

Theorem 887. *For $f \ge 0$ and every $t > 0$, $\mathrm{Ent}_\mu(f) \le \int f \log f \,\mathrm{d}\mu - (\int f \,\mathrm{d}\mu) \log t - \int f \,\mathrm{d}\mu + t$; for a probability measure this is $\mathrm{Ent}_\mu(f) \le \int (f \log \frac{f}{t} - f + t) \,\mathrm{d}\mu$, with equality for $t = \int f \,\mathrm{d}\mu$.*

Proof. ☐

Theorem 888. *For a probability measure μ, $f \geq 0$ and h with $f,\ f\log f,\ fh,\ e^h \in L^1(\mu)$, $\int fh\,\mathrm{d}\mu \leq \mathrm{Ent}_\mu(f) + (\int f\,\mathrm{d}\mu)(\int e^h\,\mathrm{d}\mu - 1)$.*

Proof. Step 1: if '∫ f dμ = 0' then 'f = 0' a.e. and both sides vanish.
Step 2: Young's inequality with 'v = h + log ∫ f dμ', integrated. ☐

Theorem 889. *For a probability measure μ and $f \geq 0$ with $f,\ f\log f \in L^1(\mu)$, $\mathrm{Ent}_\mu(f) \geq 0$.*

Proof. ☐

Theorem 890. *Let μ, ν be probability measures with $\mathrm{d}\nu = h\,\mathrm{d}\mu$, $0 \leq h \leq K$. Then $\mathrm{Ent}_\nu(f) \leq K\ \mathrm{Ent}_\mu(f)$ for every $f \geq 0$ with $f,\ f\log f \in L^1(\mu) \cap L^1(\nu)$.*

Proof. With $t = \int f\,\mathrm{d}\mu > 0$ and $\varphi_t(u) = u\log u - u\log t - u + t \geq 0$, $\mathrm{Ent}_\nu(f) \leq \int \varphi_t(f)\,\mathrm{d}\nu = \int h\,\varphi_t(f)\,\mathrm{d}\mu \leq K\int \varphi_t(f)\,\mathrm{d}\mu = K\ \mathrm{Ent}_\mu(f)$. If $\int f\,\mathrm{d}\mu = 0$ both entropies vanish. ☐

Theorem 891. *Let $\mathrm{d}\nu = h\,\mathrm{d}\mu$ with $0 < c \leq h \leq K$. Then $\int \psi\,\mathrm{d}\mu \leq c^{-1}\int \psi\,\mathrm{d}\nu$ for every integrable $\psi \geq 0$.*

Proof. ☐

Theorem 892. (F3, Holley–Stroock.) *If μ satisfies $\mathrm{LSI}(\alpha)$, ψ is measurable with $m \leq \psi \leq M$ and $\nu := Z^{-1}e^{-\psi}\mu$, then ν satisfies $\mathrm{LSI}(\alpha e^{-(M-m)})$.*

Proof. The density $h = e^{-\psi}/Z$ of ν satisfies $e^{-M}/Z \leq h \leq e^{-m}/Z$. By Lemma 890, $\mathrm{Ent}_\nu(g^2) \leq \frac{e^{-m}}{Z}\,\mathrm{Ent}_\mu(g^2) \leq \frac{e^{-m}}{Z}\frac{2}{\alpha}\int|\nabla g|^2\,\mathrm{d}\mu$, and by Lemma 891, $\int|\nabla g|^2\,\mathrm{d}\mu \leq Ze^M\int|\nabla g|^2\,\mathrm{d}\nu$; the normalization cancels and the constants multiply to e^{M-m}. ☐

Theorem 893. *If g and T are differentiable and $\|DT(x)\| \leq L$ for all x, then $|\nabla(g\circ T)(x)| \leq L\,|\nabla g(T(x))|$.*

Proof. ☐

Theorem 894. (F4, transport.) *If μ satisfies $\mathrm{LSI}(\alpha)$ and T is a proper C^1 map with $\|DT\| \leq L$, then $T_\#\mu$ satisfies $\mathrm{LSI}(\alpha/L^2)$.*

Proof. For $g \in C_c^1$ on the target, $g\circ T \in C_c^1$ and $|\nabla(g\circ T)| \leq L\,|\nabla g|\circ T$ (Lemma 893), so $\mathrm{Ent}_{T_\#\mu}(g^2) = \mathrm{Ent}_\mu((g\circ T)^2) \leq \frac{2}{\alpha}\int|\nabla(g\circ T)|^2\,\mathrm{d}\mu \leq \frac{2L^2}{\alpha}\int|\nabla g|^2\,\mathrm{d}T_\#\mu$. ☐

Theorem 895. (F4, transport.) *If μ satisfies $\mathrm{LSI}(\alpha)$ and T is a proper C^1 map which is Lipschitz with constant L, then $T_\#\mu$ satisfies $\mathrm{LSI}(\alpha/L^2)$.*

Proof. ☐

Theorem 896. *For probability measures μ, ν and a bounded measurable $f \geq 0$ on $X \times Y$,*

$$\mathrm{Ent}_{\mu\otimes\nu}(f) \leq \int \mathrm{Ent}_\mu(f(\cdot,y))\,\mathrm{d}\nu(y) + \int \mathrm{Ent}_\nu(f(x,\cdot))\,\mathrm{d}\mu(x).$$

Proof. With $F(y) = \int f(x,y)\,\mathrm{d}\mu(x)$ and $a = \int f\,\mathrm{d}\mu\otimes\nu$, Fubini gives $\mathrm{Ent}_{\mu\otimes\nu}(f) = \int \mathrm{Ent}_\mu(f(\cdot,y))\,\mathrm{d}\nu(y) + \mathrm{Ent}_\nu(F)$. For $M > 0$ let $h = \log\max\{F/a, e^{-M}\}$, so that $\int e^h\,\mathrm{d}\nu \leq 1 + e^{-M}$ and $\mathrm{Ent}_\nu(F) \leq \int h\,\mathrm{d}\nu = \iint f(x,y)h(y)\,\mathrm{d}\nu(y)\,\mathrm{d}\mu(x)$. By the variational inequality $\int f(x,\cdot)h\,\mathrm{d}\nu \leq \mathrm{Ent}_\nu(f(x,\cdot)) + e^{-M}\int f(x,\cdot)\,\mathrm{d}\nu$ (Lemma 888), hence $\mathrm{Ent}_\nu(F) \leq \int \mathrm{Ent}_\nu(f(x,\cdot))\,\mathrm{d}\mu(x) + ae^{-M}$; let $M \to \infty$. ☐

Theorem 897. *For a differentiable g on the Euclidean product $E_1 \times E_2$, $|\nabla g(x,y)|^2 = |\nabla_x g(\cdot,y)(x)|^2 + |\nabla_y g(x,\cdot)(y)|^2$.*

Proof. □

Theorem 898. (F2, tensorization.) *If μ_1 satisfies $\mathrm{LSI}(\alpha_1)$ on E_1 and μ_2 satisfies $\mathrm{LSI}(\alpha_2)$ on E_2, then $\mu_1 \otimes \mu_2$ satisfies $\mathrm{LSI}(\min\{\alpha_1,\alpha_2\})$ on the Euclidean product $E_1 \times E_2$.*

Proof. For $g \in C_c^1(E_1 \times E_2)$, Lemma 896 with $f = g^2$, the LSI of μ_1 and μ_2 applied to the slices $g(\cdot,y)$, $g(x,\cdot)$, Fubini and Lemma 897 give $\mathrm{Ent}_{\mu_1\otimes\mu_2}(g^2) \le \frac{2}{\alpha_1}\int|\nabla_x g|^2 + \frac{2}{\alpha_2}\int|\nabla_y g|^2 \le \frac{2}{\min\{\alpha_1,\alpha_2\}}\int|\nabla g|^2\,\mathrm{d}\mu_1 \otimes \mu_2$. □

Theorem 899. (F2, tensorization.) *If μ_1 satisfies $\mathrm{LSI}(\alpha_1)$ and μ_2 satisfies $\mathrm{LSI}(\alpha_2)$, then $\mu_1 \otimes \mu_2$ satisfies $\mathrm{LSI}(\min\{\alpha_1,\alpha_2\})$ on $\Theta = \mathbb{R} \times Z$.*

Proof. □

Definition 900. For $R > 0$ let $\chi_R(x) := \phi(2 - |x|^2/R^2)$, where $\phi \in C^\infty(\mathbb{R})$ is nondecreasing with $\phi = 0$ on $(-\infty, 0]$ and $\phi = 1$ on $[1,\infty)$. Then $0 \le \chi_R \le 1$, $\chi_R = 1$ on $\{|x| \le R\}$, $\chi_R = 0$ on $\{|x| \ge \sqrt{2}R\}$ and $|\nabla\chi_R| \le C/R$ with $C = 4\sup|\phi'|$.

Theorem 901. *There is $C \ge 0$ such that $|\nabla\chi_R(x)| \le C/R$ for all $R > 0$ and x.*

Proof. $\nabla\chi_R(x) = -\phi'(2 - |x|^2/R^2)\,2x/R^2$ vanishes for $|x| > 2R$ (where the argument of ϕ' is negative) and is bounded by $\sup|\phi'| \cdot 4R/R^2$ otherwise. □

Theorem 902. *Let μ satisfy $\mathrm{LSI}(\alpha)$ on a finite-dimensional space, and let $\rho = p\,\mu$ be a probability measure with $p > 0$ of class C^1, $\nabla \log p \in L^2(\rho)$ and $\mathrm{KL}(\rho\|\mu) < \infty$. Then $\mathrm{KL}(\rho\|\mu) \le \frac{1}{2\alpha}\int|\nabla\log p|^2\,\mathrm{d}\rho$.*

Proof. Apply the LSI to $g_n = \chi_n\sqrt{p}$ with the cutoffs $\chi_n = \chi_{n+1}$ of Definition 900. By dominated convergence ($|u\log u| \le 1 + |p\log p|$ for $0 \le u \le p$, and $p\log p \in L^1(\mu)$ since $\mathrm{KL}(\rho\|\mu) < \infty$), $\mathrm{Ent}_\mu(g_n^2) \to \int p\log p\,\mathrm{d}\mu = \mathrm{KL}(\rho\|\mu)$. Since $\nabla g_n = \chi_n\nabla\sqrt{p} + \sqrt{p}\,\nabla\chi_n$ and $|\nabla\chi_n| \le C/(n+1)$ (Lemma 901), $|\nabla g_n|^2 \le |\nabla\sqrt{p}|^2 + \frac{C}{n+1}(p + |\nabla\sqrt{p}|^2) + \frac{C^2}{(n+1)^2}p$, whose integral tends to $\int|\nabla\sqrt{p}|^2\,\mathrm{d}\mu = \frac{1}{4}\int|\nabla\log p|^2\,\mathrm{d}\rho$. □

Theorem 903. (L1.) *If μ satisfies $\mathrm{LSI}(\alpha)$ on a finite-dimensional space and $\rho \ll \mu$ is a probability measure with a smooth density and $\mathrm{KL}(\rho\|\mu) < \infty$, then $\mathrm{KL}(\rho\|\mu) \le \frac{1}{2\alpha}I(\rho|\mu)$.*

Proof. Lemma 902 applied to the density defining $I(\rho|\mu)$. □

Theorem 904. *Let ρ, ρ', μ be probability measures with $\mathrm{KL}(\rho\|\mu) < \infty$ and $\mathrm{KL}(\rho'\|\mu) < \infty$, and let $\varepsilon \in [0,1]$. Then $\mathrm{KL}((1-\varepsilon)\rho + \varepsilon\rho'\|\mu) \le (1-\varepsilon)\,\mathrm{KL}(\rho\|\mu) + \varepsilon\,\mathrm{KL}(\rho'\|\mu) < \infty$. (Proof: $\frac{\mathrm{d}\rho_\varepsilon}{\mathrm{d}\mu} = (1-\varepsilon)\frac{\mathrm{d}\rho}{\mathrm{d}\mu} + \varepsilon\frac{\mathrm{d}\rho'}{\mathrm{d}\mu}$ and $\mathrm{KL}(\rho\|\mu) = \int\phi(\frac{\mathrm{d}\rho}{\mathrm{d}\mu})\,\mathrm{d}\mu$ with $\phi(x) = x\log x + 1 - x$ convex.)*

Proof. Step 1: the density of the mixture is the mixture of the densities.
Step 2: pointwise convexity of ‘klFun‘.
Step 3: integrability of both sides.
Step 4: integrate. □

Theorem 905. *For $\nu = w\,\nu_0$ with $w > 0$, $\mathrm{KL}(\nu\|\nu_0) = \int(w\log w + 1 - w)\,\mathrm{d}\nu_0 \le \int(w-1)^2\,\mathrm{d}\nu_0 = \chi^2(\nu\|\nu_0)$, since $\log w \le w - 1$.*

Proof. □

Theorem 906. *For $v > 0$ and $t \geq 0$, $\mathcal{N}(0, v)(\{|x| > t\}) = 2\bar{\Phi}(t/\sqrt{v}) \leq e^{-t^2/(2v)}$. (For $x \geq t \geq 0$, $x^2 \geq t^2 + (x-t)^2$, so $\phi_v(x) \leq e^{-t^2/(2v)}\phi_v(x-t)$; integrate over $x > t$ and use $\mathcal{N}(0, v)(0, \infty) \leq \frac{1}{2}$.)*

Proof. ☐

Theorem 907. *For $v > 0$ and $t \geq 0$, $\int_t^\infty x\, \mathcal{N}(0, v)(\,\mathrm{d}x) = \sqrt{v/(2\pi)}\, e^{-t^2/(2v)} = v\, \phi_v(t)$.*

Proof. ☐

Theorem 908. *Let $v > 0$, $|m| \leq m_0 \leq T$ and $X \sim \mathcal{N}(m, v)$. Then $\mathbb{E}[|X|\, 1_{\{|X|>T\}}] \leq (m_0 + \sqrt{2v/\pi})e^{-(T-m_0)^2/(2v)}$. (Write $X = m + W$ with $W \sim \mathcal{N}(0, v)$: $|X| > T$ forces $|W| > t := T - m_0$ and $|X| \leq m_0 + |W|$, so the left side is at most $m_0\mathbb{P}(|W| > t) + \mathbb{E}[|W|1_{\{|W|>t\}}]$, and Lemmas 906 and 907 bound the two terms.)*

Proof. ☐

Theorem 909. *Let $K \geq 0$ and $\Phi\colon [0, \infty) \to \mathbb{R}$ with $\Phi(t)\,(1 + K(t - s)) \leq \Phi(s)$ for all $0 \leq s \leq t$. Then $\Phi(t) \leq e^{-Kt}\Phi(0)$ for all $t \geq 0$. (Iterating on the partition kt/n gives $\Phi(t)(1 + Kt/n)^n \leq \Phi(0)$; let $n \to \infty$.)*

Proof. Step 1: iterate on the partition ‘k t / n‘: ‘Φ(k t/n) (1 + K t/n)^k ≤ Φ(0)‘.
Step 2: ‘Φ t ≤ Φ 0 / (1 + K t / n)^n‘ for every ‘n ≥ 1‘.
Step 3: let ‘n → ∞‘, using ‘(1 + x/n)^n → e^x‘. ☐

Theorem 910. *For $\rho^2 + c^2 = 1$ and every $x \in \mathbb{R}$, $\int He_n(\rho x + cz)\,\gamma(\,\mathrm{d}z) = \rho^n He_n(x)$: the Hermite polynomials are the eigenfunctions of the Ornstein–Uhlenbeck semigroup. (Two-step induction with the recursion $He_{n+2} = XHe_{n+1} - (n + 1)He_n$ and Gaussian integration by parts.)*

Proof. ☐

Definition 911. For $\rho \in [-1, 1]$, the law of $(X, \rho X + \sqrt{1 - \rho^2}Z)$ with X, Z independent standard Gaussians: the standard Gaussian pair with correlation ρ.

Theorem 912. *Both marginals of the correlated standard Gaussian pair are γ.*

Proof. ☐

Theorem 913. *For $|\rho| \leq 1$ and standard Gaussians (X, Y) with correlation ρ, $\mathbb{E}[He_m(X)He_n(Y)] = \rho^n\, n!\, \delta_{mn}$. (Proof: $\mathbb{E}[He_n(\rho x + \sqrt{1 - \rho^2}Z)] = \rho^n He_n(x)$ by the three-term recursion and Gaussian integration by parts, then orthogonality.)*

Proof. ☐

Theorem 914. *If $g \in L^2(\gamma)$ and $\int gHe_n\, \mathrm{d}\gamma = 0$ for all $n \geq 0$, then $g = 0$ γ-a.e.: the Hermite polynomials are complete in $L^2(\gamma)$. (Every monomial is a combination of Hermite polynomials, so all moments of $g\gamma$ vanish; by dominated convergence its Fourier transform vanishes, and the positive and negative parts of $g\gamma$ have the same characteristic function, hence coincide.)*

Proof. ☐

Definition 915. $a_n(f) := \frac{1}{\sqrt{n!}} \int f\, He_n\, \mathrm{d}\gamma$, the coefficient of $f \in L^2(\gamma)$ on the normalized Hermite polynomial $He_n/\sqrt{n!}$.

Definition 916. The normalized Hermite polynomials $He_n/\sqrt{n!}$, $n \geq 0$, form an orthonormal basis of $L^2(\gamma)$ (orthogonality and completeness).

Theorem 917. *For $f \in L^2(\gamma)$, $\|f\|^2_{L^2(\gamma)} = \sum_{n\geq 0} a_n(f)^2$, the series converging.*

Proof. $\square$

Theorem 918. *For $F, G \in L^2(\gamma)$ and $|\rho| \leq 1$, $\mathbb{E}[F(X)G(Y)] = \sum_{n\geq 0}\langle He_n/\sqrt{n!}, F\rangle\langle He_n/\sqrt{n!}, G\rangle\rho^n$. (Expand F and G in the Hermite basis, use $\mathbb{E}[He_m(X)He_n(Y)] = \rho^n\, n!\, \delta_{mn}$ and the continuity of $(F, G) \mapsto \mathbb{E}[F(X)G(Y)]$ on $L^2(\gamma) \times L^2(\gamma)$.)*

Proof. $\square$

Theorem 919. *For $f, g \in L^2(\gamma)$, $|\rho| \leq 1$ and standard Gaussians (X, Y) with correlation ρ, $\mathbb{E}[f(X)g(Y)] = \sum_{n\geq 0} a_n(f)\, a_n(g)\, \rho^n$, the series converging (absolutely, by Parseval and $|\rho| \leq 1$).*

Proof. $\square$