Practice midterm — solutions

Problem 1. CONDITIONING AS PROJECTION

(a) A random vector X in $\mathbb{R}^d$ is Gaussian with mean μ and covariance Σ if its characteristic function is the Gaussian one:

$$\mathbb{E}e^{-i\langle X,\xi\rangle} = \exp\Big(-i\langle\mu,\xi\rangle - \tfrac{1}{2}\xi^\top\Sigma\xi\Big) \qquad \text{for every } \xi\in\mathbb{R}^d,$$

where $\mu\in\mathbb{R}^d$ and Σ is symmetric and positive semidefinite. The point of defining it this way, rather than by a density, is that it makes sense when Σ is singular — which is exactly the situation in part (c), where the vector $(X-AY, Y)$ may well be degenerate.

(b) The theorem says two things: $X-AY$ is independent of Y, and $\mathbb{E}[X\mid Y]=AY$. The first is the real content; the second follows from it in one line.

(c) The whole proof is one covariance computation plus the fact we are allowed to quote.

First, $(X-AY,\,Y)$ is a linear image of (X,Y) — multiply by $\begin{pmatrix} I & -A \\ 0 & I\end{pmatrix}$ — so it is again a centered Gaussian vector. That is what lets us talk about independence through covariances at all. Now compute the cross-covariance:

$$\mathrm{Cov}(X-AY,\,Y) = \Sigma_{XY} - A\Sigma_{YY} = \Sigma_{XY} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YY} = 0.$$

This is of course how A was chosen: it is the unique matrix that kills the cross-covariance. Jointly Gaussian and uncorrelated means independent, so $X-AY$ is independent of Y.

For the conditional expectation, write $X = (X-AY)+AY$. The second piece is a function of Y, so conditioning on Y leaves it alone; the first piece is independent of Y, so conditioning on Y does nothing and leaves its plain expectation, which is 0. Hence $\mathbb{E}[X\mid Y] = 0 + AY$.

It is worth saying what the theorem is geometrically: AY is the orthogonal projection of X onto the span of Y (in the L^2 sense), and $X-AY$ is the residual. For general random variables the residual is merely uncorrelated with Y; for Gaussians it is independent. That is the one place the Gaussian assumption is used, and it is used exactly once.

(d) Signal plus noise is the theorem with $X=w$ and $Y=y$. The pair (w,y) is a linear image of (w,ξ), so it is a centered Gaussian vector, and the three blocks we need are

$$\Sigma_{ww} = \tau^2 I, \qquad \Sigma_{wy} = \mathrm{Cov}(w,\,w+\sigma\xi) = \tau^2 I, \qquad \Sigma_{yy} = \tau^2 I + \sigma^2 I.$$

(The cross block is $\tau^2 I$, not $(\tau^2+\sigma^2)I$: the noise is independent of w, so it contributes nothing to $\mathrm{Cov}(w,y)$.) Then $A = \Sigma_{wy}\Sigma_{yy}^{-1} = \frac{\tau^2}{\tau^2+\sigma^2} I$, and the theorem gives

$$\mathbb{E}[w\mid y] = \frac{\tau^2}{\tau^2+\sigma^2}\,y, \qquad \mathrm{Cov}(w\mid y) = \Sigma_{ww} - A\Sigma_{yw} = \tau^2 I - \frac{\tau^4}{\tau^2+\sigma^2} I = \frac{\tau^2\sigma^2}{\tau^2+\sigma^2}\,I.$$

So the posterior mean is the observation shrunk toward 0 by the factor $\tau^2/(\tau^2+\sigma^2)$, and the posterior covariance is a scalar that does not depend on what was observed. This is ridge regression in its simplest form: $H=I$ and $\lambda=\sigma^2/\tau^2$.

Problem 2. Lindeberg's theorem

This is the swapping argument from the i.i.d. central limit theorem, with one change: there are no third moments, so the cubic remainder bound must be replaced by the quadratic one wherever a summand is large. Lindeberg's condition says exactly that the large summands do not matter.

No single term matters. First, Lindeberg's condition forces every individual variance to be small. Split the second moment of $Y_{n,i}$ at ε:

$$\sigma_{n,i}^2 = \mathbb{E}\big[Y_{n,i}^2 \mathbf{1}\{|Y_{n,i}| \le \varepsilon\}\big] + \mathbb{E}\big[Y_{n,i}^2 \mathbf{1}\{|Y_{n,i}| > \varepsilon\}\big] \le \varepsilon^2 + L_n(\varepsilon).$$

The right side does not depend on i, so letting $n \to \infty$ and then $\varepsilon \to 0$ gives $\max_i \sigma_{n,i} \to 0$. This is needed at the end, on the Gaussian side.

The swap. Fix a test function $f \in C^3$ with f'' and f''' bounded. Let $g_{n,i} \sim \mathcal{N}(0, \sigma_{n,i}^2)$ be independent of each other and of the array. Their sum is $\mathcal{N}(0,1)$, because the variances add up to 1, so $\mathbb{E}f(g) = \mathbb{E}f(\sum_i g_{n,i})$: we compare T_n with a Gaussian sum having the same variance profile, term by term, and replace the Y's by g's one at a time. Writing

$$W_{n,i} = g_{n,1} + \cdots + g_{n,i-1} + Y_{n,i+1} + \cdots + Y_{n,k_n}$$

for everything except the ith slot, the sum with $i-1$ swaps is $W_{n,i} + Y_{n,i}$ and the sum with i swaps is $W_{n,i} + g_{n,i}$. Telescoping,

$$\big|\mathbb{E}f(T_n) - \mathbb{E}f(g)\big| \le \sum_{i=1}^{k_n} \Big|\mathbb{E}f\big(W_{n,i} + Y_{n,i}\big) - \mathbb{E}f\big(W_{n,i} + g_{n,i}\big)\Big|.$$

The point is that $W_{n,i}$ is independent of both $Y_{n,i}$ and $g_{n,i}$. Taylor-expand both arguments around $W_{n,i}$. The constant terms agree. The linear terms $\mathbb{E}[Y_{n,i} f'(W_{n,i})]$ and $\mathbb{E}[g_{n,i} f'(W_{n,i})]$ factor by independence into a mean times something, and both means are 0. The quadratic terms factor into $\sigma_{n,i}^2 \, \mathbb{E}f''(W_{n,i})$ on both sides, since $Y_{n,i}$ and $g_{n,i}$ have the same variance. So everything cancels except the remainders:

$$\big|\mathbb{E}f(W_{n,i} + Y_{n,i}) - \mathbb{E}f(W_{n,i} + g_{n,i})\big| \le \mathbb{E}\big|R(Y_{n,i})\big| + \mathbb{E}\big|R(g_{n,i})\big|.$$

The remainders. This is the step that differs from the i.i.d. case. On the Y side nothing is assumed about third moments, so use the cubic bound only where $|Y_{n,i}| \le \varepsilon$, where it reads $|Y|^3 \le \varepsilon Y^2$, and the quadratic bound where $|Y_{n,i}| > \varepsilon$:

$$\sum_i \mathbb{E}|R(Y_{n,i})| \le \frac{\|f'''\|_\infty}{6} \, \varepsilon \sum_i \mathbb{E}Y_{n,i}^2 + \|f''\|_\infty \sum_i \mathbb{E}\big[Y_{n,i}^2 \mathbf{1}\{|Y_{n,i}| > \varepsilon\}\big] = \frac{\|f'''\|_\infty}{6} \, \varepsilon + \|f''\|_\infty \, L_n(\varepsilon).$$

On the Gaussian side there is nothing to truncate, so use the cubic bound throughout: $\mathbb{E}|g_{n,i}|^3 = \sigma_{n,i}^3 \mathbb{E}|g|^3$ and $\sum_i \sigma_{n,i}^3 \le \max_i \sigma_{n,i} \cdot \sum_i \sigma_{n,i}^2 = \max_i \sigma_{n,i}$, so

$$\sum_i \mathbb{E}|R(g_{n,i})| \le \frac{\|f'''\|_\infty \, \mathbb{E}|g|^3}{6} \max_i \sigma_{n,i},$$

which is where the first step is needed.

Conclusion. Altogether $|\mathbb{E}f(T_n) - \mathbb{E}f(g)| \le \frac{\|f'''\|_\infty}{6}\varepsilon + \|f''\|_\infty L_n(\varepsilon) + \frac{\|f'''\|_\infty \mathbb{E}|g|^3}{6} \max_i \sigma_{n,i}$. Let $n \to \infty$: the second term vanishes by Lindeberg's condition and the third by the first step, leaving $\limsup_n |\mathbb{E}f(T_n) - \mathbb{E}f(g)| \le \frac{\|f'''\|_\infty}{6}\varepsilon$. Since ε was arbitrary, the limit is 0. This holds for every smooth test function, and the reduction lemma upgrades that to convergence in distribution.

Problem 3. the norm of a Gaussian vector

(a)(i) An ε-net of (T,d) is a subset $T' \subset T$ such that every point of T is within distance ε of some point of T'. The covering number $N(T,d,\varepsilon)$ is the smallest size of an ε-net; it is $+\infty$ if no finite net exists.

(ii) For a centered Gaussian process $(X_t)_{t\in T}$, the canonical metric is the L^2 distance between the coordinates,

$$d(s,t) = \big(\mathbb{E}(X_s - X_t)^2\big)^{1/2} = \sqrt{\Sigma_{s,s} - 2\Sigma_{s,t} + \Sigma_{t,t}},$$

where $\Sigma_{s,t} = \mathbb{E}[X_s X_t]$ is the covariance kernel. It is the metric in which the process "sees" its index set.

(b) Any finite collection $(\langle g,u_1\rangle,\dots,\langle g,u_n\rangle)$ is a linear image of g, so it is a centered Gaussian vector: X is a centered Gaussian process. Its supremum is explicit, which is what makes this example a good test case: by Cauchy–Schwarz $\langle g,u\rangle \le \|g\|$ for every unit u, with equality at $u = g/\|g\|$, so $\sup_u X_u = \|g\|$. The canonical metric is just as explicit: $X_u - X_v = \langle g, u-v\rangle$ is Gaussian with variance $\|u-v\|^2$, so $d(u,v) = \|u-v\|$ — the Euclidean metric — and the diameter is $D = 2$, attained by antipodal points.

So the question "how big is $\mathbb{E}\|g\|$" is the same as "how big is the maximum of a Gaussian process indexed by the sphere," and the two theorems answer it from the geometry of the sphere alone.

(c) Dudley's integral runs to $D/2 = 1$. Inserting the given covering bound,

$$\mathbb{E}\|g\| \le 12\int_0^1 \sqrt{\log N(S^{d-1},\|\cdot\|,\varepsilon)}\,d\varepsilon \le 12\int_0^1 \sqrt{d\,\log(1+2/\varepsilon)}\,d\varepsilon = C\sqrt{d},$$

with $C = 12\int_0^1 \sqrt{\log(1+2/\varepsilon)}\,d\varepsilon$. The dimension comes out of the square root and the integral that remains is a number (about 1.35, so $C \approx 16$).

(d) Sudakov's inequality holds at every scale, and one scale is enough. At $\varepsilon = \frac14$ the given lower bound says $\log N \ge (d-2)\log 2$, so

$$\mathbb{E}\|g\| \ge c\cdot\tfrac14\sqrt{(d-2)\log 2} \ge c'\sqrt{d} \qquad \text{for } d \ge 4,$$

using $d - 2 \ge d/2$. Together with (c): $c'\sqrt{d} \le \mathbb{E}\|g\| \le C\sqrt{d}$. The sphere in dimension d has exponentially many well-separated points, and that is the only property of it the two theorems used.

The constants are far from sharp — the truth is $\sqrt{d} - 1 \le \mathbb{E}\|g\| \le \sqrt{d}$, which one gets from the second and fourth moments of $\|g\|$ — but the order $\sqrt{d}$ is right, and it was obtained without computing a single moment.