Module 4M24: Computational Statistics and Machine Learning

4M24 Tripos 2022/23 - Cribs

1. (a) i. $\mathsf{Dev}(\hat{\mathcal{I}}) = \mathbb{E}\{(\hat{\mathcal{I}} - \mathcal{I})^2\} = \mathbb{E}\{\sum_{m,n} w_m w_n f(x_m) f(x_n)\} - 2\mathcal{I}\mathbb{E}\{\sum_n w_n f(x_n)\} + \mathcal{I}^2$ which in matrix form is $\mathcal{I}^2\left[\mathbf{w}^\mathsf{T}\mathbf{C}\mathbf{w} - 2\mathbf{w}^\mathsf{T}\mathbf{1} + 1\right]$ if $C_{ij} = 1, \;\; i \neq j$ and $C_{ij} = \mathbb{E}\{f^2(X)\}/\mathcal{I}^2, \;\; i = j$
Marks 30 available - 10 for component form of variance and 20 for correctly defining both diagnonal and off-diagonal components of the matrix

ii. $\partial_w \mathsf{Dev}(\hat{\mathcal{I}}) = 2\mathcal{I}^2\mathbf{C}\mathbf{w} - 2\mathcal{I}^2\mathbf{1}$ so $\mathbf{w} = \mathbf{C}^{-1}\mathbf{1}$
Marks 10 available - 5 for derivative and 5 for correct expression

(b) $\mathbf{C} = \epsilon\mathbf{I} + \mathbf{1}\mathbf{1}^\mathsf{T}$ with $\epsilon = \frac{\mathbb{E}\{f^2(X)\}}{\mathcal{I}^2} - 1$ using Sherman-Morrison $\mathbf{C}^{-1} = \frac{1}{\epsilon}(\mathbf{I} - \frac{1}{\epsilon+N}\mathbf{1}\mathbf{1}^\mathsf{T})$ then $\mathbf{w} = \mathbf{C}^{-1}\mathbf{1} = \frac{1}{\epsilon}(\mathbf{I} - \frac{1}{\epsilon+N}\mathbf{1}\mathbf{1}^\mathsf{T})\mathbf{1} = \frac{1}{\epsilon+N}\mathbf{1}$ so each $w_n = \frac{1}{\epsilon+N}$ and $\frac{1}{\epsilon+N}\sum_{n=1}^N f(x_n)$ follows.
Marks 40 available - 10 for correct definition of C, 10 for correct inverse, 10 noting that all weights are equal and correct definition, 5 for correct final expression, and 5 to state that estimator impractical as ϵ requires knowledge of object that is being estimated $\mathcal{I}$

(c) i. If unbiased $\mathbb{E}\{\hat{\mathcal{I}}\} = \mathcal{I}$ so $\mathbb{E}\{\hat{\mathcal{I}}\} = \frac{1}{\epsilon+N}\sum_{n=1}^N \mathbb{E}\{f(x_n)\} = \frac{N}{\epsilon+N}\mathcal{I}$. therefore estimate is biased multiplicatively by $\frac{N}{\epsilon+N}$
Marks 10 available - 10 available showing multiplicative bias

ii. As $N \to \infty$ then $\frac{N}{\epsilon+N}\mathcal{I} \to \mathcal{I}$ so bias is consistently reduced as $N \to \infty$.
Marks 10 available - 10 showing that estimate though biased is consistent

2. (a) i. Require to define the transition kernel operator $P(x, dy)$ so that it has $\pi(\cdot)$ as its invariant density. Consider some function $p(x, y)$ and define a transition operator as

$$P(x, dy) = p(x, y)dy + r(x)\delta_x(dy)$$

If $p(x, x) = 0$ and $\int_{\mathcal{R}^d} P(x, dy) = 1$ then $r(x) = 1 - \int_{\mathcal{R}^d} p(x, y)dy$, probability that chain remains at x. If $p(x, y)$ satisfies reversibility $\pi(x)p(x, y) = \pi(y)p(y, x)$ then $\pi(\cdot)$ is the invariant density of the transition kernel $P(x, \cdot)$
Marks 15 available - 10 for definition of conditions and 5 for form of $r(x)$

ii. $\int_{\mathcal{R}^d} P(x, A)\pi(x)dx$ is equal to

$$\begin{aligned}
& \int \left[\int_A p(x, y)dy\right] \pi(x)dx + \int r(x)\delta_x(A)\pi(x)dx \\
= & \int_A \left[\int p(x, y)\pi(x)dx\right] dy + \int_A r(x)\pi(x)dx \\
= & \int_A \left[\int p(y, x)\pi(y)dx\right] dy + \int_A r(x)\pi(x)dx \\
= & \int_A (1 - r(y))\pi(y)dy + \int_A r(x)\pi(x)dx \\
= & \int_A \pi(y)dy = \pi^*(A)
\end{aligned}$$

The Reversibility condition for $p(x, y)$ is sufficient in designing a transition operator to target a specific invariant distribution Marks 25 available for coherent layout of proof

(b) i. Now need to define the form of $p(x, y)$ explicitly. Consider a Proposal density $q(x, y)$ s.t. $\int q(x, y)dy = 1$, if $q(x, y)$ is reversible we are finished. This may hold $\pi(x)q(x, y) > \pi(y)q(y, x)$ x to y transitions too frequent. Rebalance both sides by introducing an Acceptance probability $\alpha(x, y)$. In this case $\alpha(x, y) < 1$ so now reversibility is established as $\pi(x)q(x, y)\alpha(x, y) = \pi(y)q(y, x)\alpha(y, x)$. Need form of $\alpha(x, y)$, maximum value of α is 1, set $\alpha(y, x) = 1$. Then $\pi(x)q(x, y)\alpha(x, y) = \pi(y)q(y, x)$ and

$$\alpha(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}$$

What about if $\pi(x)q(x, y) < \pi(y)q(y, x)$? $\alpha(x, y) = 1$ Marks 25 available for correctly defining and writing out the form of the acceptance probability

ii. We now have that $p_{MH}(x,y) = q(x,y)\alpha(x,y)$ for $x \neq y$ with

$$\alpha(x,y) = min\left(\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right)$$

Overall transition kernel $P_{MH}(x,dy)$ is equal to

$$q(x,y)\alpha(x,y)dy + \left[1 - \int_{\mathcal{R}^d} q(x,y)\alpha(x,y)dy\right]\delta_x(dy)$$

is reversible by construction and hence has $\pi(x)$ as its invariant density. Marks 15 available -for correctly describing the overall form of the transition kernel

(c) If $q(x,y)$ is symmetric i.e. $q(x,y) = q(y,x)$ e.g. $\mathcal{N}(x|y,\Sigma) = \mathcal{N}(y|x,\Sigma)$ then

$$\alpha(x,y) = min\left(\frac{\phi(y)}{\phi(x)}, 1\right)$$

ratio of unnormalised densities only.

for $j = 1 \rightarrow N$ **do**
  Simulate y from $q(x^{(j)}, \cdot)$
  Simulate u from $U(0,1)$
  if $u \leq \alpha(x^{(j)}, y)$ **then**
    Set $x^{(j+1)} = y$
  else
    Set $x^{(j+1)} = x^{(j)}$
  end if
end for
Return $\{x^{(1)}, x^{(2)}, x^{(3)}, \cdots, x^{(N)}\}$

Marks 20 available - for clearly written pseudo-code noting that ratio of proposals is one

3. (a) RKHS is a Hilbert space of functions where point evaluation is continuous. In L2 two functions f and g can be close in L2 norm but not so in pointwise manner. The smaller Hilbert space, the RKHS, is such that the evaluation of functions is continuous that is if functions are close in norm they are also close pointwise 15 marks for correctly defining continuous nature of function evaluation pointwise

(b) The RKHS is defined by the reproducing kernel function and likewise a reproducing kernel function defines an RKHS. Kernel functions are both symmetric and positive definite. The reproducing property describes $f(x) =< f(.), K(.,x) >$. 20 marks in total - 5 marks for first two and 10 marks for reprodcing property

(c) i. If the regularised log-likelihood function is

$$\sum_{i=1}^{N} \log(p(y_i|\mathbf{x}_i)) - \frac{\lambda}{2}\|\log\hat{\mu}\|^2$$

the gradient takes the form $\mathbf{K}(\mathbf{y} - \hat{\boldsymbol{\mu}} - \lambda\boldsymbol{\alpha})$ where $\mathbf{K}$ where each $K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$ and vector $\hat{\boldsymbol{\mu}}$ with elements $\exp(\sum_{n=1}^{N}\alpha_n k(\mathbf{x}_i, \mathbf{x}_n))$
30 marks for gradient and definition of matrix and vector

ii. The matrix of second derivatives is $-\mathbf{KVK} - \lambda\mathbf{K}$ with matrix $\mathbf{V}$ defined as $diag(\hat{\boldsymbol{\mu}})$
20 marks for Hessian and definition of matrix V

iii. Correctly plugging the above into the Newton iteration should yield
$\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} + (\mathbf{KVK} + \lambda\mathbf{K})^{-1}\mathbf{K}(\mathbf{y} - \hat{\boldsymbol{\mu}} - \lambda\boldsymbol{\alpha})$
Marks 15 available

4. (a) $\log p(x) \propto -\frac{\nu+1}{2}\log\left(1 + \frac{x^2}{\nu}\right)$ so $\frac{\partial}{\partial x}\log p(x) = -\frac{\nu+1}{2} \times \frac{2x}{\nu+x^2} = -\frac{(\nu+1)x}{\nu+x^2}$ so Langevin diffusion is $dx = -\frac{(\nu+1)x}{\nu+x^2}dt + \sqrt{2}dW$ where W is Brownian motion.
Marks 40 available - 20 for derivative and 20 for correct SDE

(b) ULA is Euler-Maruyama discrete version of Langevin diffusion $x_{k+1} = x_k - \tau\frac{(\nu+1)x_k}{\nu+x_k^2} + \sqrt{2\tau}z_k$ were τ is integration step size and z_k is zero-mean, unit variance Gaussian variable.
Marks 20 available - 10 for Euler-Mayuma discretisation and 10 for correct noise scaling definition

(c) MALA proposal from ULA is $q(x'|x) \propto \exp\left(-\frac{1}{4\tau}|x' - x + \tau\frac{(\nu+1)x}{\nu+x^2}|^2\right)$ and reverse is $q(x|x') \propto \exp\left(-\frac{1}{4\tau}|x - x' + \tau\frac{(\nu+1)x'}{\nu+x'^2}|\right.$

Ratio of $\log\frac{p(x')}{p(x)} = \frac{-\frac{\nu+1}{2}\log\left(1+\frac{x'^2}{\nu}\right)}{-\frac{\nu+1}{2}\log\left(1+\frac{x^2}{\nu}\right)}$ so $\frac{p(x')}{p(x)} = \left[\frac{\nu+x^2}{\nu+x'^2}\right]^{\frac{\nu+1}{2}}$.

So $\alpha(x', x) = \min\left(1, \left[\frac{\nu+x^2}{\nu+x'^2}\right]^{\frac{\nu+1}{2}} \times \frac{q(x|x')}{q(x'|x)}\right)$

Marks 40 available - 10 for each correct forward and backward proposals (20 in total) and 20 for correctly and compactly defining the acceptance probability