

Midterm practice

About the midterm exam

- You may bring one “cheat sheet”. This is one piece of paper you may refer to during the quiz where you can write anything you want. No other materials such as notes, books, computer, phone, networked watch, etc. are allowed. They must be put away and inaccessible during the quiz.
- The midterm will cover everything we have covered up to now, including material tested in the quiz. Use the practice quiz questions for examples of midterm questions that test the earlier material. The questions below are only for material covered after the quiz, but the midterm will cover both.
- Write all answers in the answer book (“blue book”) you will be given. The question sheet will not be handed in or graded.
- Points will be deducted for anything you write that is wrong even if you also give a correct answer. Cross out anything you think is wrong.
- Guessing is discouraged. You will get more points for not answering a question than for giving a wrong answer.
- The midterm will be 60 minutes in recitation on October 21.
- Explain your answers with a few words. True/false answers without explanation may get no points. Calculations without explanations, even if correct, may not get full credit. For True/False questions: if it’s true, say why in a few words (not a formal proof). A mathematical statement is false if there’s a counterexample. For example, “All prime numbers are odd” is false because 2 is a counterexample. If the statement is “All prime numbers are odd”, you can answer: “no, 2 is prime and even”. You probably would get full credit for saying: “no, 2”
- There will be True/False questions that should be quick to answer and full answer questions that involve reasoning or calculations.
- The practice questions below are to let you know the kind of questions that might be asked. The actual midterm will may be different, but in a similar spirit.

True/False

1. IF $(X, Y) \in \mathbb{R}^2$ is a two component random variable and both X and Y are Gaussian, then (X, Y) is a jointly Gaussian pair.
2. “The” maximum likelihood estimator $\hat{\theta}(X)$ is unique.
3. If you compute the value of a function $f(x)$ in double precision arithmetic with machine precision $\epsilon_{\text{mach}} < 10^{-16}$, then the computed value will be accurate to about 16 decimal places.
4. In order to do maximum likelihood parameter estimation, you must have a probability model that determines the probability density of an observation, X , as a function of the parameter set θ .

5. The singular values of a matrix A determine the eigenvalues of $A^T A$.
6. It is impossible to produce (computationally) samples of a random variable $X \sim f(\cdot)$ without using a formula for $f(x)$.
7. It is impossible to reproduce exactly the results of a Monte Carlo simulation using (pseudo) random numbers because you get different random numbers every time you run a program.
8. If A is a square matrix, not necessarily symmetric, and $Ax = \lambda x$ with $x \neq 0$ and λ real, then if $x^T y = 0$ then $x^T A y = 0$ (if yes, why? if no, counter-example).
9. If T is an exponential random variable with $\bar{T} = E[T]$, then, for any t_0 , no matter how large, $E[T - t_0 | T > t_0] = \bar{T}$.
10. If A is a rectangular matrix then $A^T A$ is positive definite.
11. If X has PDF $f(x) = \frac{1}{Z} \frac{1}{1+x^2}$ (a Cauchy distribution centered at zero), then the sample means $\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$ converge to 0 as $n \rightarrow \infty$ ($X_k \sim f$, i.i.d.).

Multiple choice questions

1. Which information about a symmetric positive definite matrix A determines the fastest possible convergence rate of gradient descent applied to the minimization problem

$$\min_x f(x) \quad , \quad f(x) = \|Ax - b\|_2 \quad .$$

- (a) The largest eigenvalue of A
 - (b) The smallest eigenvalue of A
 - (c) The 2-norm condition number of A
 - (d) $\|b\|_2$
2. If $\theta = (\theta_1, \theta_2)$ is a two component parameter vector and $\hat{\theta}$ is the maximum likelihood estimate using dataset $(X_1, \dots, X_n)$. Which is the most natural 95% confidence region in θ space
 - (a) The box in parameter space defined by $[\hat{\theta}_1 - a_1, \hat{\theta}_1 + a_1]$ and $[\hat{\theta}_2 - a_2, \hat{\theta}_2 + a_2]$, these being confidence intervals for θ_1 and θ_2 separately.
 - (b) A disk in parameter space defined by $\|\theta - \hat{\theta}\|_2 \leq r$.
 - (c) The ellipsoid in parameter space $(\theta - \hat{\theta})^T H (\theta - \hat{\theta}) < r$, with H being the Fisher information matrix.
 - (d) The ellipsoid in parameter space $(\theta - \hat{\theta})^T C (\theta - \hat{\theta}) < r$, with C being covariance matrix which is the inverse of the Fisher information matrix.

Full answer questions

1. Suppose A is a large $n \times d$ matrix whose entries are unknown, but you have a routine that can compute Ax for any $x \in \mathbb{R}^d$ and $A^T y$ for any $y \in \mathbb{R}^n$. Explain an iterative algorithm to estimate the largest singular value of A using these operations many times. *Hint*, singular values of A are eigenvalues of $A^T A$ and of AA^T , you choose which.
2. Find a formula for $\nabla f(x)$, where $f(x) = \log(\|Ax - b\|_2)$.

3. We want to find θ so that

$$\mathbb{E} \left[\frac{1}{X - \theta} \right] = 1 .$$

Suppose $X \sim f(\cdot)$, but f is unknown. Instead we have n independent samples $X_k \sim f$. We estimate θ by finding $\hat{\theta}$ with

$$\frac{1}{n} \sum_k \frac{1}{X_k - \hat{\theta}} = 1 .$$

Describe a way to estimate $\text{var}(\hat{\theta})$ using the same dataset. Show that $\hat{\theta}$ is approximately normal for large n if the central limit theorem applies (tails are not too big).

4. Suppose X is multi-variate normal with mean zero and covariance matrix C with information matrix $H = CD^{-1}$. Find a formula for

$$\frac{1}{dr} \Pr(\|X^T H X\|_2 < r^2) .$$

Hint. Express X in principle components $X = u_1 Y_1 + u_2 Y_2$ (u_1 and u_2 are eigenvectors of H). Then Y_1 and Y_2 are independent and $\|X^T H X\|_2 < r$ is equivalent to $Z_1^2 + Z_2^2 < r^2$, where $Z_1 = a_1 Y_1$ and $Z_2 = a_2 Y_2$ with appropriate a_1 and a_2 so that Z_1 and Z_2 are independent standard normals. Then (polar coordinates)

$$\Pr(Z_1^2 + Z_2^2 \leq r^2) = \frac{1}{2\pi} \int_{s=0}^r 2\pi s e^{-\frac{s^2}{2}} ds .$$

5. Suppose T_k are independent samples and we want to fit to a *gamma* distribution

$$f(t) = \frac{1}{Z_{\lambda,p}} t^p e^{-\lambda t} .$$

Describe how to estimate p and λ using maximum likelihood estimation. Write a formula for the likelihood and/or log likelihood function that should be maximized. The algorithm may use a software function that evaluates

$$\Gamma(x) = \int_0^\infty s^{x-1} e^{-s} ds .$$

This is the traditional definition of the *Gamma function*, but

$$\int_0^\infty s^p e^{-s} ds = \Gamma(p+1) .$$

You can see how Z depends on λ using the substitution $\lambda t = s$ in the normalization integral for f that defines Z .

6. Find an algorithm that samples the PDF

$$f(x) = \frac{1}{Z} \begin{cases} x & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

The sampler should take $X = \phi(U)$, where U is a standard uniform as given by a random number generator.