ORF 544
Stochastic Optimization and Learning

Midterm exam
Spring, 2019

---

a) This is an open book, take-home midterm. It is due at 4pm, Monday, April 1. Turn the exam in to Kim Lupinacci in Room 117 in Sherrerd Hall.

b) Under no circumstances are you to discuss the midterm with *anyone*.

c) The questions below take you on a tour through the course textbook *Stochastic Optimization and Learning*. Each question is numbered xx.yy, where xx indicates the chapter from which the question is derived.

1.1 (5 points) What are the three classes of state variables? Explain the differences.

1.2 (5 points) What are the two strategies for designing policies for sequential decision problems? Give the basic equation for each.

2.1 (15 points) In chapter 9, we present the *universal objective function* in equation (9.20). Show how the following problems can be written using this form:

a) The final reward formulation of the newsvendor problem.

b) The cumulative reward formulation of the newsvendor problem.

c) The asymptotic formulation of the newsvendor problem. What are the differences between the asymptotic formulation and the final reward formulation?

2.2 (5 points) What is the state variable for the multi-armed bandit problem as described in section 2.1.9?

2.3 (10 points) We wish to use Q-learning to solve the problem of deciding whether to continue playing a game where you win $\$1$ if you flip a coin and see heads, and lose $\$1$ if you see tails. Assume that this is a stationary (infinite horizon) problem with a discount factor of $\gamma = 0.9$. Using a stepsize $\alpha = .5$, execute three iterations of the Q-learning algorithm in equations (2.19) and (2.21). Initialize your estimates $\bar{Q}(s,a) = 0$.

3.1 (10 points) Show that $\mathbb{E}\left[\left(\bar{\mu}^{n-1} - \mu(n)\right)^2\right] = \lambda^{n-1}\sigma^2 + (\beta^{n-1})^2$ (equation (3.24) in chapter **??**).

3.2 (10 points) Use equations (3.16) and (3.17) to update the mean vector with prior

$$\mu^0 = \begin{bmatrix} 10 \\ 18 \\ 12 \end{bmatrix}.$$

Assume that we observe $W = 19$ after observing $x = 3$ and that our prior covariance matrix Σ^0 is given by

$$\Sigma^0 = \begin{bmatrix} 12 & 4 & 2 \\ 4 & 8 & 3 \\ 2 & 3 & 10 \end{bmatrix}.$$

Assume that $\lambda^W = 4$. Give μ^1 and Σ^1.

3.3 (10 points) In this exercise you will use the equations in section 3.8.1 to update a linear model. Assume you have an estimate of a linear model given by

$$\begin{aligned} \bar{F}(x|\theta^0) &= \theta_0 + \theta_1\phi_1(x) + \theta_2\phi_2(x) \\ &= -20 + 3.1\phi_1 + 5.6\phi_2. \end{aligned}$$

Assume that the matrix B^0 is a 3×3 identity matrix. Assume the vector $\phi = (\phi_0 \quad \phi_1 \quad \phi_2) = (1 \quad 27 \quad 12)$ and that you observe $\hat{f}^1 = 140$. Give the updated regression vector θ^1.

4.1 (5 points) Describe what is meant by the sample average approximation method. Is this a deterministic or stochastic optimization problem? Explain.

4.2 (5 points) Give a bound on the performance of the solution from using a sample average approximation when applied to a convex optimization problem (the formula is in the book - you do not need to derive the bound). Why is this bound important, given that it involves an unknown coefficient (which means we have no idea how tight the bound is)?

4.3 (5 points) What is the difference between a stationary policy, a deterministic nonstationary policy, and an adaptive policy?

5.1 (5 points) Use a stochastic gradient algorithm to solve the problem

$$\min_x \frac{1}{2}(X - x)^2,$$

where X is a random variable. Use a harmonic stepsize rule (equation (6.15)) with parameter $\theta = 5$. Perform three iterations assuming that you observe $X^1 = 6, X^2 = 2, X^3 = 5$. Use a starting initial value of $x^0 = 10$. What is the best possible formula for θ for this problem? (The answer to this is contained in chapter 6).

5.2 (10 points) Write out the objective function for finding the best stepsize policy for a derivative-based algorithm to solve the general problem $\max_x \mathbb{E}_W\{F(x, W)|S_0\}$ using:

a) The final reward.

b) The cumulative reward.

For each case, write the expectation indicating what random variable is involved (as we did writing $\mathbb{E}_W$ to indicate that we are taking the expectation over W).

5.3 (10 points) Consider solving a concave stochastic optimization problem using a stochastic gradient algorithm using Kesten's stepsize policy (see section 6.2.3).

a) Formulate this as a dynamic program by specifying the state variables, the decision variable, the exogenous information (what is learned after a decision is made), the transition function (give the equations showing how each state variable changes), and the objective function (assume you have a finite computation budget).

b) What choices have to be made in designing a stepsize policy?

c) It seems reasonable that the best stepsize rule depends on the starting position x^0. Assuming that this is true, how does this change the optimization problem in (a), given that each time you solve the problem, you might want to choose a different starting point.

6.1 (5 points) We would like to solve the problem

$$\min_\theta \mathbb{E}\frac{1}{2}(\theta - W)^2$$

using a stochastic gradient algorithm. Show that if we use a stepsize rule $\alpha_{n-1} = 1/n$, then θ^n is a simple average of observations of W given by $W^1, W^2, \ldots, W^n$ (thus proving equation 6.14).

6.2 (10 points) We again would like to solve the problem

$$\min_{\theta} \mathbb{E}\frac{1}{2}(\theta - W)^2$$

using a stochastic gradient algorithm. This time assume that the observations W come from the model

$$W^n = 5 + 10(1 - e^{-\rho n}) + \varepsilon^n,$$

where $\varepsilon \sim N(0, \sigma^2)$.

a) Given the objective function that the BAKF stepsize formula optimizes. What random variables are involved in the expectation?

b) Show that if $\rho = 0$ (that is, the series is stationary) then the BAKF stepsize returns the stepsize $\alpha_{n-1} = \frac{1}{n}$.

c) Show that if $\sigma^2 = 0$ then $\alpha_{n-1} = 1$.

6.3 (5 points) What is the difference between deterministic and stochastic stepsize policies?

7.1 Consider the problem of finding the best in a set of discrete choices $\mathcal{X} = \{x_1, \ldots, x_M\}$. Assume that for each alternative you maintain a lookup table belief model, where $\bar{\mu}_x^n$ is your estimate of the true mean μ_x, with precision β_x^n. Assume that your belief about μ_x is Gaussian, and let $X^\pi(S^n)$ be a policy that specifies the experiment $x^n = X^\pi(S^n)$ that you will run next, where you will learn $W_{x^n}^{n+1}$ which you will use to update your beliefs.

a) (10 points) Formulate this learning problem as a stochastic optimization problem. Define your state variable, decision variable, exogenous information, transition function and objective function.

b) (5 points) Specify three possible policies, with no two from the same policy class (PFA, CFA, VFA and DLA).

7.2 You have three alternatives, with priors (mean and precision) as given in the first line of the table below. You then observe each of the alternatives in three successive experiments, with outcomes shown in the table. All observations are made with precision $\beta^W = 0.2$. Assume that beliefs are independent.

Iteration	A	B	C
Prior $(\bar{\mu}_x^0, \beta_x^0)$	(12,0.2)	(18,0.2)	(8,0.2)
1	-	-	10
2	15	-	-
3	-	17	-

Table 21.1 Three observations, for three alternatives, given a normally distributed belief, and assuming normally distributed observations.

a) (5 points) Give the objective function (algebraically) for offline learning (maximizing final reward) if you have a budget of three measurements, and where you evaluate the policy using the truth (as you would do in a simulator).

b) (5 points) Give the numerical value of the policy that was used to generate the choices that created table 21.1, using our ability to use the simulated truth. Assume that $\mu = \bar{\mu}^0$ for the simulated truth. This requires minimal calculations (which can be done without a calculator).

c) (5 points) Now assume that you need to do these measurements in an online (cumulative reward) setting. Give the objective function (algebraically) to find the optimal policy for online learning (maximizing cumulative reward) if you have three measurements. Using the numbers in the table, give the performance of the policy that generated the choices that were made. (This again requires minimal calculations.)

7.3 (15 points) There are seven alternatives with normally distributed priors on μ_x for $x \in \{1, 2, 3, 4, 5, 6, 7\}$ given in the table below: Without doing any calculations,

Choice	$\bar{\mu}^n$	σ^n
1	5.0	9.0
2	3.0	8.0
3	5.0	10.0
4	4.5	12.0
5	5.0	8.0
6	5.5	6.0
7	4.0	8.0

Table 21.2 Priors

state any relationships between the alternatives based on the knowledge gradient. For example, $1 < 2 < 3$ means 3 has a higher knowledge gradient than 2 which is better than 1 (if this was the case, you do not have to separately say that $1 < 3$).

7.4 Assume that we have a standard normal prior about a true parameter μ which we assume is normally distributed with mean μ^0 and variance $(\sigma^0)^2$.

a) (5 points) Given the observations $W^1, \ldots, W^n$, is $\bar{\mu}^n$ deterministic or random?

b) (5 points) Given the observations $W^1, \ldots, W^n$, what is $\mathbb{E}(\mu|W^1, \ldots, W^n)$ (where μ is our truth)? Why is μ random given the first n measurements?

c) (5 points) Given the observations $W^1, \ldots, W^n$, what is the mean and variance of $\bar{\mu}^{n+1}$? Why is $\bar{\mu}^{n+1}$ random?

7.5 (30 points) You have to find the best of five alternatives. After n measurements, you have the data given in the table below. Assume that the precision of the measurement is $\beta^W = 0.8$.

a) (10 points) Give the definition of the knowledge gradient, first in plain English and second using mathematics.

Z	Cumulative normal	Normal density	Z	Cumulative normal	Normal density
-4	0.000032	0.000134	0	0.500000	0.398942
-3.9	0.000048	0.000199	0.1	0.539828	0.396953
-3.8	0.000072	0.000292	0.2	0.579260	0.391043
-3.7	0.000108	0.000425	0.3	0.617911	0.381388
-3.6	0.000159	0.000612	0.4	0.655422	0.368270
-3.5	0.000233	0.000873	0.5	0.691462	0.352065
-3.4	0.000337	0.001232	0.6	0.725747	0.333225
-3.3	0.000483	0.001723	0.7	0.758036	0.312254
-3.2	0.000687	0.002384	0.8	0.788145	0.289692
-3.1	0.000968	0.003267	0.9	0.815940	0.266085
-3	0.001350	0.004432	1	0.841345	0.241971
-2.9	0.001866	0.005953	1.1	0.864334	0.217852
-2.8	0.002555	0.007915	1.2	0.884930	0.194186
-2.7	0.003467	0.010421	1.3	0.903200	0.171369
-2.6	0.004661	0.013583	1.4	0.919243	0.149728
-2.5	0.006210	0.017528	1.5	0.933193	0.129518
-2.4	0.008198	0.022395	1.6	0.945201	0.110921
-2.3	0.010724	0.028327	1.7	0.955435	0.094049
-2.2	0.013903	0.035475	1.8	0.964070	0.078950
-2.1	0.017864	0.043984	1.9	0.971283	0.065616
-2	0.022750	0.053991	2	0.977250	0.053991
-1.9	0.028717	0.065616	2.1	0.982136	0.043984
-1.8	0.035930	0.078950	2.2	0.986097	0.035475
-1.7	0.044565	0.094049	2.3	0.989276	0.028327
-1.6	0.054799	0.110921	2.4	0.991802	0.022395
-1.5	0.066807	0.129518	2.5	0.993790	0.017528
-1.4	0.080757	0.149728	2.6	0.995339	0.013583
-1.3	0.096800	0.171369	2.7	0.996533	0.010421
-1.2	0.115070	0.194186	2.8	0.997445	0.007915
-1.1	0.135666	0.217852	2.9	0.998134	0.005953
-1	0.158655	0.241971	3	0.998650	0.004432
-0.9	0.184060	0.266085	3.1	0.999032	0.003267
-0.8	0.211855	0.289692	3.2	0.999313	0.002384
-0.7	0.241964	0.312254	3.3	0.999517	0.001723
-0.6	0.274253	0.333225	3.4	0.999663	0.001232
-0.5	0.308538	0.352065	3.5	0.999767	0.000873
-0.4	0.344578	0.368270	3.6	0.999841	0.000612
-0.3	0.382089	0.381388	3.7	0.999892	0.000425
-0.2	0.420740	0.391043	3.8	0.999928	0.000292
-0.1	0.460172	0.396953	3.9	0.999952	0.000199
0	0.500000	0.398942	4	0.999968	0.000134

Figure 21.1 Cumulative standard normal distribution and normal density. Do not interpolate - simply use the closest value of Z.

Choice	θ^n	β^n	β^{n+1}	$\tilde{\sigma}$	$\max_{x' \neq x} \theta^n_{x'}$	ζ	$f(\zeta)$	ν^{KG}_x
1	12.0	0.06	0.86	3.9375	15	-0.762	0.128	0.504
2	13.0	0.04	?	?	?	?	?	?
3	15.0	0.05	0.85	4.3386	13	-0.461	0.210	0.911
4	10.0	0.01	0.81	9.9381	15	-0.503	0.197	1.958
5	8.0	0.0625	0.8625	3.8523	15	-1.817	0.014	0.054

b) (10 points) Fill in the missing entries for alternative 3. Be sure to clearly write out each expression and then perform the calculation. For the knowledge gradient ν^{KG}_x, you will need to use the table on the next page. Do not interpolate; round the Z value to the nearest 0.1 and use the corresponding entry.

c) (10 points) Now assume that we have an online learning problem. We have a budget of 20 measurements, and the data in the table above shows what we have learned after five measurements. Assuming no discounting, what is the online knowledge gradient for alternative 4 (I have intentionally chosen an alternative for which the offline knowledge gradient is given in the table)? Give both the formula and the number.