Algorithmical and Statistical Modelling, Fall 2007, Midterm

Note: the points of the problems below sum to 120, of which a maximum of 100 will be scored for the course grade ("safety margin").

Problem 1. (30 points, challenge: display mathematical rigour) Let $(\Omega, \mathcal{F}, P)$ be a *probability space* and $A, B \in \mathcal{F}$ two *events*. Prove that if $P(A \triangle B) = 0$, then $P(A) = P(B)$.
Hints: The *symmetric difference* of two sets A and B is defined by

$$A \triangle B := (A \cup B) \setminus (A \cap B) = (A \cup B) \cap (A \cap B)^{\mathrm{c}} = (A \setminus B) \cup (B \setminus A)\,.$$

According to the lecture notes, a *probability space* is a triple $(\Omega, \mathcal{F}, P)$ (of a nonvoid set Ω, a nonvoid family of subsets of Ω and a set function $P : \mathcal{F} \to [0, 1]$) satisfying the following *Kolmogorov axioms*:

(K1) $\Omega \in \mathcal{F}$;
(K2) $A \in \mathcal{F}$ implies $A^{\mathrm{c}} \in \mathcal{F}$ (closure under complement);
(K3) $\{A_n\}_{n=1,2,\cdots} \subseteq \mathcal{F}$ implies $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$ (closure under countable union);
(K4) $P(\Omega) = 1$;
(K5) for every pairwise disjoint sequence $(A_n)_{n=1,2,\cdots}$ in $\mathcal{F}$ it holds that

$$P(\biguplus_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n) \qquad (\sigma\text{-additivity})\,,$$

where the symbol $\uplus$ indicates "disjoint union".

Here are three elementary probabilities of probability spaces, which follow from the Kolmogorov axioms:

(P1) $\varnothing \in \mathcal{F}$;
(P2) $P(\varnothing) = 0$;
(P3) if $A, B \in \mathcal{F}$, then also $A \triangle B \in \mathcal{F}$;
(P4) if $A, B \in \mathcal{F}$, then also $A \cap B \in \mathcal{F}$.

You can directly use the above (K1–K5), (P1– P4) and basic facts from set theory such as $A \subseteq A \cup B$ in your proof. Beyond that, you should prove **everything** *you say, and also make sure that every item that you introduce*

is well-defined.

Proof: For any $A, B \in \mathcal{F}$, it follows from (K2, P4) that $A \setminus B = A \cap B^{\mathrm{c}} \in \mathcal{F}$, i.e.,

(P5) $A, B \in \mathcal{F}$ implies $A \setminus B \in \mathcal{F}$ — and so $P(A \setminus B)$ has definition.

Furthermore, if $A, B \in \mathcal{F}$ are disjoint, putting $A_1 = A$, $A_2 = B$ and $A_n = \varnothing$ for $n \geqslant 3$, by (K5, P2) we get $P(A \uplus B) = P(A) + P(B)$. Thus, if $A, B \in \mathcal{F}$ and $A \subseteq B$, then $P(B) = P(A \uplus (B \setminus A)) = P(A) + P(B \setminus A) \geqslant P(A)$. Summarizing, we have

(P6) $A, B \in \mathcal{F}$ and $A \subseteq B$ imply $P(B \setminus A) = P(B) - P(A)$ and $P(A) \leqslant P(B)$.

Since $A \cap B \subseteq A \cup B$, by the definition of $A \triangle B$ and (P6), we have $P(A \cup B) - P(A \cap B) = P(A \triangle B) = 0$, i.e., $P(A \cup B) = P(A \cap B)$. Since $A \cap B \subseteq A \subseteq A \cup B$, we know $P(A \cap B) \leqslant P(A) \leqslant P(A \cup B)$. It is now clear that $P(A) = P(A \cap B)$. Similarly, $P(B) = P(A \cap B)$. Thus, $P(A) = P(B)$.

Problem 2. (20 points, challenge: conceptual design.) In sciences that are concerned with historical texts (like religion, history of literature, history, and others), one is sometimes confronted with the question whether two texts have been written by the same or by two different authors. This can refer to texts that come in different documents, or to sections within a longer document (like the books that make up the Bible, or certain large portions within a the *Niebelungenlied*, a classical German mideval verse saga). Your task: describe a probabilistic model which text historians could use as the basis for the task of deciding whether two texts have a common author. Provide the following components of such a model: (i) (5 points) an underlying event space Ω, described in plain English, (ii) (15 points) suitable random variables, where each RV is specified through its observation space E – specify these as rigorously as possible. Justify why you propose *these* random variables as a basis for the author decision task. Take into consideration that two texts may not only be from two different authors, but also typically are about two different topics...

Note: as always in nontrivial modelling cases, there is not a unique correct solution but considerable freedom of design.

Target length of your answer: about the same as the length of this problem statement.

Solution. (i) A possible Ω would be the set of all events ω where some author at some time in the past produces some text. (ii) For RVs, a simplistic but not stupid approach is to introduce two different types of RVs. The first, X, yields authors, and is a hidden variable: $X(\omega)$ is the author of the textwriting event ω; the observation space is the finite set of authors that lived on earth so far. The second type of RV, Y_i, where $i = 1, 2, ...$ gives the length of the i-th sentence in a text, i.e. $Y_i(\omega)$ is the i-th word in the text written in the textwriting event ω. For mathematical "total rigour", one could define $Y_i(\omega) = \epsilon$ in case that i is larger than the text has sentences. The observation space for the variables Y_i is $\mathbb{N}$. One may (even must) assume that the Y_i are identically distributed (and to some degree, at least for i, i' which are sufficiently spaced, also independently). The rationale behind this choice is that different authors have different styles, and an important (although not uniquely identifying) aspect of style is the length distribution of sentences. If the empirical sentence length distributions from two texts differs significantly (by some statistical test tailored to this situation), one may this, with due caution, as an indication that the two texts have different authors. It would sharpen the test if additional style-characterizing variables were measured.

Problem 3. (20 points; challenge: understanding acceptance functions) (a.) (5 points) Show that the Boltzmann acceptance function,

$$A(\mathbf{x}^* \mid \mathbf{x}) = \frac{g(\mathbf{x}^*)}{g(\mathbf{x}^*) + g(\mathbf{x})}, \tag{1}$$

where g is the pdf of the distribution from which is being sampled, $\mathbf{x}^*$ is the proposed new sample point offered by the proposal distribution, and $\mathbf{x}$ is the current sample point, has the detailed balance property, when the proposal distribution $S(\mathbf{x}^* \mid \mathbf{x})$ is symmetric. (Notice that here we refer to a "global" version of these distributions, not to "local" ones that update dimensions individually, as in the lecture notes).

(b.) (5 points) Why would you think that the Boltzmann acceptance function is used more rarely than the Metropolis acceptance function?

(c.) (10 points) Find some other acceptance function which gives detailed balance in conjunction with a symmetric proposal distribution, and which is different from either the Boltzmann or the Metropolis acceptance function (prove detailed balance!).

Solution. (a.) We have to show that $g(\mathbf{x})A(\mathbf{x}^* \mid \mathbf{x})S(\mathbf{x}^* \mid \mathbf{x}) = g(\mathbf{x}^*)A(\mathbf{x} \mid \mathbf{x}^*)S(\mathbf{x} \mid \mathbf{x}^*)$. This follows immediately from (1) by an elementary 1-step

transformation, using symmetry of S.

(b.) One reason may be the close ties that the Metropolis acceptance function enjoys with statistical physics, which makes it so nicely interpretable as an energy game and which also connects it to simulated annealing. Another (likely more important) reason is that the Boltzmann acceptance function always yields a lower accepting rate than the Metropolis acceptance function, and is thus more expensive.

(c.) The easiest way to obtain a new proposal distribution is by linear mixing from the Boltzmann and Metropolis distributions. Details are straightforward and omitted here.

Problem 4. (20 points, challenge: formalization of real-world items for simulated annealing.) This is a simulated annealing design task. Each semester, the registrar's office has to create a schedule for the courses, assigning a time and a room to each course. There are numerous side-conditions, some strict (e.g., two courses must not be scheduled to the same room at the same time), some only "gradual" (e.g., professor A prefers not to have courses at 14:15 because he usually fetches his kids from school around that time). Finding a good schedule that observes all strict constraints and a close-to-optimal negotiation between the softer constraints is a demanding combinatorial optimization problem, and simulated annealing is one way to approach it. Your task: sketch out some essentials of how to set up a simulated annealing optimization of this task. Specifically, give an account of the following points.

1. (7 points) What are the microstates in this task? Give a formal specification, introducing relevant variables in words. You may assume that the schedule is weekly-periodic.

2. (7 points) The cost function here will be designed as a sum of various components, each of which takes care of one constraint. One of these might be, *it is desirable to distribute the sessions of one course rather evenly over the week, not lumping them closely together.* Start from the formalism that you introduced to describe microstates, and formalize a sum term in the cost function that would take care of this constraint.

3. (6 points) One way to ensure that the strict constraints are never violated is to assign an infinite cost to them. This is theoretically possible, but dangerous; if infinite cost terms are used, greatest care has to be used when defining the proposal distribution. Why?

Solution. 1. Let S be the set of weekly sessions that have to be scheduled (that is, for each course that has n weekly sessions, S contains n elements), T be the set of time slots over the entire week, and L the set of possible locations. All of these are finite sets. A microstate is a mapping $s : S \to T \times L$, where it is a matter of convention whether one requires that s is injective or not (non-injective s clearly would make useless schedules, but might be handy in simulated annealing as states that can be used to "bridge" different areas of the state space, ensuring ergodicity).

2. One way to go: for the sessions of one course, introduce a penalty that is inversely proportional to the minimum time lag between two sessions in a week. To make this formal, let C be the set of courses, and $c : S \to C$ the labelling of sessions by the courses, and $t : S \to T$ the assignment of sessions to times. Let minInterval be the function that assigns to any finite, nonempty set of times in a week the shortest time lag between two times in the set if the set of times contains at least two times, and the duration of an entire week if the set contains only one time. Then, a cost function term that captures the soft constraint might be

$$\sum_{x \in C} \frac{1}{\text{minInterval}(t(c^{-1}(x)))}.$$

3. The danger lies in the fact that states to which infinite costs are assigned can never be reached by the simulated annealing state search; they are insurmountable energy barriers. If they are used, the design of the proposal function must ensure that between every pair of finite-cost states there is a sequence of other finite-cost states, generateble by iterated proposals, to ensure ergodicity. This may be hard to verify.

Problem 5. (30 points, challenge: clear mathematical understanding of simulated annealing basics) Assume that simulated annealing is used to find the lowest-energy state in a system that has (only) three states s_1, s_2, s_3, where $E(s_1) = 0, E(s_2) = 1, E(s_3) = -\infty$. The search is started from s_1. The proposal distribution is given by the following transition matrix $S_{ij} = S(s_j \mid s_i)$:

$$S(s_j \mid s_i) = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 \end{pmatrix}.$$

a. (15 points) Give the transition kernel $T(s_j \mid s_i)$ (at some fixed temperature T) for the simulated annealing process which results from this setup.

Concretely, provide a 3×3 transition probability matrix. **b.** (15 points) Give a non-trivial upper bound for the probability to reach state s_3 within n steps, after a start in s_1, at a temperature T.

Solution. a. First observation: at all transitions ij where $S_{ij} = 0$, also $T(s_j \mid s_i) = 0$. The $s_1 \to s_2$ transition is "energetically uphill", so the transition probability $T(s_2 \mid s_1)$ is the product of the proposal probability (which is $1/2$) with the acceptance probability, which is $\exp(\Delta E/T) = \exp(-1/T)$. The reverse direction transition probability $T(s_1 \mid s_2)$ is equal to the proposal probability, i.e. is $1/2$, because this transition is "energetically downhill" and acceptance is certain. The probability to transit from s_3 to s_2 is zero because the jump would be "energetically infinitely far uphill". Observing that rows in the matrix of $T(s_j \mid s_i)$ must sum to 1, we get from these findings

$$T(s_j \mid s_i) = \begin{pmatrix} 1 - \frac{1}{2}\exp(-1/T) & \frac{1}{2}\exp(-1/T) & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 0 & 1 \end{pmatrix}.$$

b. The precise probability $P(1, 3, n)$ to reach s_3 from s_1 within n steps, at a temperature T, would be $T^n(1, 3)$, where T is the transition matrix from **a.**. However, this is not easily written in closed form. A relevant upper bound can however be given by observing that $P(1, 3, n) \leq 1 - P(\text{within } n \text{ steps, } s_2 \text{ is never entered from } s_1)$. This leads to a bound

$$P(1, 3, n) \leq 1 - (1 - \frac{1}{2}\exp(-1/T))^{n-1}.$$