Learning to Match Distributions
Preparatory Homework

Romain Lopez

September 14, 2026

This is the preparatory homework for the course, and it is worth **10% of your final grade**. Its purpose is to put everyone on common ground before the technical lectures begin: every exercise below reappears later in the course, so working through it now means meeting each idea for the second time rather than the first. A well-prepared student should spend under an hour on each of the five exercises. Spending longer on a single exercise to make up some background (e.g., convex optimization) is perfectly fine. Being unable to complete most of the homework, on the other hand, is a signal that this course may not be a good match.

Due Monday, September 21, at the start of class. Submit a single PDF on Brightspace. You may typeset in LaTeX or scan legible handwritten work.

Grading policy. Each part is worth the points marked beside it, 100 in total, and those points are awarded for *a genuine attempt* rather than for a correct answer (i.e., a blank or a one-line non-answer earns nothing). Solutions are released after the deadline.

Collaboration and AI. You may discuss these problems with other students, and you may use AI assistants. Disclose both — a line at the top naming who you worked with and how you used any assistant is enough. Nothing prevents you from having one write your answers, but the preparation is the entire point, and you would be skipping it. Looking up a definition from convex analysis or multivariate calculus is ordinary and expected; outsourcing the reasoning is not.

Exercise 1 (Least squares recovers the conditional mean (25 points))**.**

Let (X, Y) be random vectors in $\mathbb{R}^p \times \mathbb{R}^q$ with joint density $p(x, y)$, and suppose the marginal $p(x) = \int p(x, y)\,\mathrm{d}y$ is strictly positive. The *conditional density* is $p(y \mid x) = p(x, y)/p(x)$, and the *conditional expectation* is

$$m(x) := \mathbb{E}[Y \mid X = x] = \int y\, p(y \mid x)\,\mathrm{d}y.$$

(a) (5 points) *Tower property.* Show that $\mathbb{E}[m(X)] = \mathbb{E}[Y]$.

(b) (5 points) *Orthogonality.* For every function $v \colon \mathbb{R}^p \to \mathbb{R}^q$, show that

$$\mathbb{E}\big[\langle Y - m(X), v(X)\rangle\big] = 0.$$

(c) (10 points) *Pythagoras.* Deduce that for every function $f \colon \mathbb{R}^p \to \mathbb{R}^q$,

$$\mathbb{E}\|Y - f(X)\|^2 = \mathbb{E}\|Y - m(X)\|^2 + \mathbb{E}\|m(X) - f(X)\|^2.$$

Conclude that m minimizes $\mathbb{E}\|Y - f(X)\|^2$ over all functions f.

(d) (5 points) *A computation.* Let (X, Y) be jointly Gaussian with means (μ_X, μ_Y) and covariance blocks $\Sigma_{XX}, \Sigma_{XY}, \Sigma_{YY}$, with Σ_{XX} invertible. Show that

$$\mathbb{E}[Y \mid X] = \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X).$$

Hint. Write $Y = \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X) + R$. Check that (X, R) is jointly Gaussian and that R is uncorrelated with X.

Exercise 2 (Pushforward and change of variables (15 points))**.**

Let X be a random vector with law μ and let T be a map. The *pushforward* $T_{\#}\mu$ is the law of $T(X)$. Equivalently, for every function φ,

$$\mathbb{E}\big[\varphi(T(X))\big] = \int \varphi(T(x))\, p(x)\, \mathrm{d}x \qquad \text{whenever } X \text{ has density } p.$$

For a differentiable map $T = (T^1, \ldots, T^d)\colon \mathbb{R}^d \to \mathbb{R}^d$, write $DT(x)$ for its *Jacobian matrix* at x: the $d \times d$ matrix whose (i, j) entry is $\partial T^i/\partial x_j(x)$.

(a) (5 points) Let $X \sim \mathcal{N}(0, I_d)$ and $T(x) = Ax + b$ with A invertible. Show that $T_{\#}\mu = \mathcal{N}(b, AA^\top)$.

(b) (5 points) Let X have density p and let T be a continuously differentiable bijection with differentiable inverse. Show that $T(X)$ has density

$$q(y) = p\big(T^{-1}(y)\big)\, \big|\det DT\big(T^{-1}(y)\big)\big|^{-1}.$$

(c) (5 points) Suppose $X = a$ with probability 1, so $\mu = \delta_a$. What is $T_{\#}\delta_a$? Show that **no** map T satisfies $T_{\#}\delta_a = \frac{1}{2}\delta_b + \frac{1}{2}\delta_c$ for $b \neq c$.

Exercise 3 (The continuity equation (20 points))**.**

For a scalar function $f\colon \mathbb{R}^d \to \mathbb{R}$, the *gradient* $\nabla f(x)$ is the vector of partial derivatives $(\partial f/\partial x_1, \ldots, \partial f/\partial x_d)$. For a vector field $u\colon \mathbb{R}^d \to \mathbb{R}^d$ with components $u = (u^1, \ldots, u^d)$, the *divergence* is the scalar

$$\nabla \cdot u(x) := \sum_{i=1}^{d} \frac{\partial u^i}{\partial x_i}(x).$$

Let $m_t \in \mathbb{R}^d$ and $\sigma_t > 0$ depend differentiably on $t \in [0, 1]$, with time derivatives written $\dot{m}_t$ and $\dot{\sigma}_t$, and let ρ_t be the density of $\mathcal{N}\big(m_t, \sigma_t^2 I_d\big)$.

(a) (10 points) *The log density.* Write $\log \rho_t(x)$ explicitly and compute $\nabla_x \log \rho_t(x)$ and $\partial_t \log \rho_t(x)$.

(b) (5 points) *The continuity equation in log form.* Let p_t be any strictly positive differentiable density and u_t any differentiable vector field. Show that

$$\partial_t p_t + \nabla \cdot (p_t u_t) = 0 \quad \iff \quad \partial_t \log p_t + \nabla \cdot u_t + \langle \nabla \log p_t, u_t \rangle = 0.$$

Hint. Apply the product rule componentwise to $\nabla \cdot (p_t u_t)$.

(c) (5 points) *Conclusion.* Using (a) and (b), verify that

$$u_t(x) = \dot{m}_t + \frac{\dot{\sigma}_t}{\sigma_t}(x - m_t)$$

satisfies the continuity equation for ρ_t.

Exercise 4 (Linear programming duality (25 points))**.**

Let $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ and $c \in \mathbb{R}^n$, and consider the linear program

$$\text{minimize} \quad c^\top x \qquad \text{subject to} \qquad Ax = b, \quad x \geq 0. \tag{LP}$$

A point $x \in \mathbb{R}^n$ is *feasible* for (LP) if $Ax = b$ and $x \geq 0$. A feasible point $x^\star$ is *optimal* if $c^\top x^\star \leq c^\top x$ for every feasible x, and $c^\top x^\star$ is then the *optimal value* of (LP).

The *Lagrangian* of (LP) is the function

$$L(x, y) := c^\top x + y^\top (b - Ax), \qquad x \geq 0, \quad y \in \mathbb{R}^m,$$

in which the equality constraint has been moved into the objective and priced by y. The *dual function* is what remains after minimizing over x,

$$g(y) := \inf_{x \geq 0} L(x, y),$$

and the *dual problem* of (LP) is

$$\text{maximize} \quad g(y) \qquad \text{over} \qquad y \in \mathbb{R}^m. \tag{D}$$

(a) (5 points) *Computing the dual function.* Show that

$$g(y) = \begin{cases} b^\top y, & \text{if } A^\top y \leq c, \\ -\infty, & \text{otherwise.} \end{cases}$$

Hint. First rewrite $L(x, y) = b^\top y + (c - A^\top y)^\top x$.

(b) (5 points) *The dual is again a linear program.* Deduce from (a) that (D) is equivalent to

$$\text{maximize} \quad b^\top y \qquad \text{subject to} \qquad A^\top y \leq c.$$

(c) (5 points) *Weak duality.* Show that $g(y) \leq c^\top x$ for every feasible x and every $y \in \mathbb{R}^m$. Conclude that the value of (D) never exceeds the optimal value of (LP).

(d) (5 points) *Matching values certify optimality.* Let $\hat{x}$ be feasible for (LP), let $\hat{y} \in \mathbb{R}^m$, and suppose the two objectives agree there,

$$c^\top \hat{x} = g(\hat{y}).$$

Show that $\hat{x}$ is then optimal for (LP), that $\hat{y}$ is optimal for (D), and that the optimal values of the two problems coincide.

(e) (5 points) *A numerical check.* Take

$$c = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \qquad A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & -1 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

and consider the two candidate points

$$\hat{x} = \left(\tfrac{1}{2},\, 0,\, \tfrac{1}{2}\right)^\top, \qquad \hat{y} = \left(\tfrac{1}{2},\, \tfrac{1}{2}\right)^\top.$$

Check that $\hat{x}$ is feasible for (LP) and that $\hat{y}$ is feasible for (D). Compute $c^\top \hat{x}$ and $b^\top \hat{y}$, and conclude from the criterion of part (d) that $\hat{x}$ and $\hat{y}$ are optimal for their respective problems.

Remark. Parts (a)–(b) identify the dual as a linear program in its own right, and part (c) shows that its value never exceeds the primal value — *weak duality.* Part (d) shows that a single pair of points with matching values closes the gap, and part (e) exhibits such a pair. That the gap always closes, for every feasible and bounded linear program, is *strong duality.*

Exercise 5 (Convexity (15 points))**.**

A function $\varphi\colon \mathbb{R}^d \to \mathbb{R}$ is *convex* if for all x, y and all $\lambda \in [0,1]$,

$$\varphi\big(\lambda x + (1-\lambda)y\big) \le \lambda\varphi(x) + (1-\lambda)\varphi(y),$$

and *strictly convex* if the inequality is strict whenever $x \neq y$ and $\lambda \in (0,1)$. Assume throughout that φ is continuously differentiable. Everything below holds verbatim with $\mathbb{R}^d$ replaced by any open convex subset of it, which is how part (c) will use it.

(a) (5 points) *First-order condition.* Show that $\varphi(y) \ge \varphi(x) + \langle \nabla\varphi(x), y - x\rangle$ for all x, y.

(b) (5 points) *Monotonicity.* Deduce that $\langle \nabla\varphi(x) - \nabla\varphi(y), x - y\rangle \ge 0$ for all x, y.

Parts (a) and (b) prove one direction of an equivalence, and the converse holds too: a continuously differentiable φ is convex *if and only if* $\langle \nabla\varphi(x) - \nabla\varphi(y), x - y\rangle \ge 0$ for all x, y, and *strictly* convex if and only if that inequality is strict whenever $x \neq y$. Use both directions freely in part (c).

(c) (5 points) *Strict convexity of the entropy.* For p in the open positive orthant $(0,\infty)^n$, define the *negative entropy*

$$h(p) := \sum_{i=1}^{n} p_i\big(\log p_i - 1\big).$$

Compute $\nabla h(p)$ and show that h is strictly convex on $(0,\infty)^n$.