

# Lecture 42: Course Review

Charles A. Bouman

Purdue University

Monday, December 7, 2026

# Outline

- 1 Inverse problems and Bayesian estimation (Ch. 1–2)
- 2 Causal Gaussian models (Ch. 3) and noncausal GMRFs (Ch. 4)
- 3 MAP with a Gaussian prior, and its optimization (Ch. 5)
- 4 Non-Gaussian MRFs (Ch. 6)
- 5 MAP with non-Gaussian priors (Ch. 7)
- 6 Surrogate functions (Ch. 8)
- 7 Constrained optimization (Ch. 9)
- 8 Plug-and-play and consensus equilibrium (Ch. 10)
- 9 The EM algorithm (Ch. 12)
- 10 Markov chains and hidden Markov models (Ch. 13)
- 11 General MRFs (Ch. 14) and stochastic simulation (Ch. 15)
- 12 Summary, and final exam logistics

# Today's Big Ideas

- The course has one core idea: model the sensor, model the image, and compute the reconstruction by optimization or sampling.
- The models grew from causal to noncausal, and from Gaussian to non-Gaussian to discrete.
- The optimization tools grew from closed-form quadratic solutions, to surrogate functions, to variable splitting, to equilibrium formulations.
- The EM algorithm estimates parameters when the measurements are incomplete.
- Markov chain theory both models data and provides the basis for the stochastic samplers.

# 1. Inverse problems and Bayesian estimation (Ch. 1–2)

- Measurements  $Y$  depend on an unknown image  $X$ ; the goal is an estimate  $\hat{X}$ . Estimating  $X$  from  $Y$  this way is called an **inverse problem**.
- **Frequentist**:  $\theta$  deterministic and unknown;  $\hat{\theta} = T(Y)$  is random;  $\text{MSE}_{\theta} = \text{var}_{\theta} + (\text{bias}_{\theta})^2$ . The **ML estimator**  $\hat{\theta} = \arg \max_{\theta} \log p_{\theta}(Y)$  is generally biased, but consistent and asymptotically efficient.
- **Bayesian**:  $X$  is random with prior  $p_x(x)$ ; minimize Bayes' risk  $\mathbb{E}[C(X, T(Y))]$ . Squared error gives MMSE  $\mathbb{E}[X|Y]$ ; absolute error gives the conditional median; zero-one gives MAP.
- **MAP**, the canonical form of the course:

$$\hat{x}_{\text{MAP}} = \arg \min_{x \in \Omega} \{ -\log p_{Y|X}(y|x) - \log p_x(x) \},$$

a balance of a **forward model** term and a **prior model** term.

- For jointly Gaussian  $X$  and  $Y$ , median = MMSE = MAP, and  $\hat{X}_{\text{MMSE}} = R_x(R_x + R_w)^{-1}Y$ .

## 2a. Causal Gaussian models (Ch. 3)

- A good predictor is a good model. Causal predictor  $\hat{X}_n = \mathbb{E}[X_n | \mathcal{F}_n]$  and error  $\mathcal{E}_n = X_n - \hat{X}_n$ ; for a zero-mean Gaussian process the predictor is linear.
- The errors are independent of the past and of each other, so in matrix form  $\mathcal{E} = (I - H)X = AX$  with  $\det(A) = 1$ , and

$$p_x(x) \propto \exp \left\{ -\frac{1}{2} x^t A^t \Lambda^{-1} A x \right\}, \quad R_X = (A^t \Lambda^{-1} A)^{-1}.$$

- **Homogeneous** models: the predictor becomes an LTI filter,  $H$  and  $A$  are Toeplitz, and inverting  $\mathcal{E}_n = X_n * (\delta_n - h_n)$  gives the white-noise-driven IIR model. For  $P < \infty$ ,  $X_n$  is an **AR process**:

$$S_X(\omega) = \sigma_c^2 / |1 - H(\omega)|^2.$$

- In 2-D, raster ordering makes  $H$  **Toeplitz block Toeplitz**, and every result carries over with 2-D convolution and the 2-D DSFT. Raster order is artificial for images and can cause artifacts.

## 2b. Noncausal Gaussian MRFs (Ch. 4)

- The goal of Chapter 4 is to remove causality from modeling. Noncausal predictor  $\hat{X}_n = \mathbb{E}[X_n|X_i \text{ for } i \neq n]$ ; the errors are no longer independent of each other, so the Chapter 3 whitening argument does not apply.
- Match  $p(x) \propto \exp\{-\frac{1}{2}x^t Bx\}$  against the conditional densities:

$$B_{i,j} = \frac{1}{\sigma_i^2} (\delta_{i-j} - g_{i,j}), \quad B = \Gamma^{-1}(I - G).$$

- A **GMRF** satisfies  $\mathbb{E}[X_n|X_i, i \neq n] = \mathbb{E}[X_n|X_i, i \in \partial n]$ , and the noncausal errors are not white:  $S_X(\omega) = \sigma_{NC}^2/(1 - G(\omega))$  — first power, where the AR result was a squared magnitude.
- **AR versus GMRF:** in 1-D a process is an order- $P$  AR model if and only if it is an order- $P$  GMRF. In 2-D an order- $P$  AR model is an order- $2P$  GMRF, and the converse fails because 2-D polynomials do not factor in general.

### 3a. MAP with a Gaussian prior (Ch. 5.1)

- Restoration model  $Y = AX + W$ ,  $A$  nonsingular  $N \times N$ ,  $W \sim N(0, \sigma^2 I)$ . Then  $\hat{x}_{ML} = A^{-1}y$ , and with  $A = U\Sigma V^t$  small singular values amplify noise; instability tracks  $\kappa(A) = |\sigma_1/\sigma_N|$ .
- With the GMRF prior  $p_x(x) \propto \exp\{-\frac{1}{2}x^t Bx\}$  the posterior is Gaussian, and  $\hat{x}_{MMSE} = \hat{x}_{MAP} = \mu(y)$  with

$$\mu(y) = \left(A^t A + \sigma^2 B\right)^{-1} A^t y, \quad R = \left(\frac{1}{\sigma^2} A^t A + B\right)^{-1}.$$

- Direct inversion of  $(A^t A + \sigma^2 B)$  is exact but impractical: one million pixels gives a  $10^{12}$ -element matrix, about 8 terabytes. Instead minimize

$$f(x) = \frac{1}{2\sigma^2} \|y - Ax\|^2 + \frac{1}{2} x^t Bx$$

numerically.

## 3b. Optimizing the quadratic MAP cost (Ch. 5.2–5.5)

- **Gradient descent:**  $x^{(k+1)} = (I - \alpha[A^t A + \sigma^2 B])x^{(k)} + \alpha A^t y$ . The error obeys  $\epsilon^{(k+1)} = (I - \alpha H)\epsilon^{(k)}$ , so with  $\alpha = 1/h_{\max}$  the slowest mode decays as  $(1 - 1/\kappa(H))^k$ . Low frequencies converge fast; **edges converge last**.
- **Line search:**  $d^{(k)} = -\sigma^2 \nabla f(x^{(k)})$ ,  $\alpha^{(k)} = \arg \min_{\alpha \geq 0} f(x^{(k)} + \alpha d^{(k)})$ ; monotone descent,  $f(x^{(k+1)}) \leq f(x^{(k)})$ .
- **ICD:**  $x_s \leftarrow \arg \min_{x_s} f(x)$ , one pixel at a time, exactly. With no blur,

$$x_s \leftarrow \frac{y_s + \lambda \sum_{r \in \partial s} g_{s-r} x_r}{1 + \lambda},$$

a weighted average of measurement and noncausal prediction.

- **Conjugate gradient** requires each new update to be conjugate to the previous ones; for quadratic  $f$  it reaches the global optimum in no more than  $N$  steps.
- All of it assumed a quadratic cost from a Gaussian prior, whose restorations are overly smooth at edges.

## 4a. Non-Gaussian MRFs: potential and influence (Ch. 6.1–6.2)

- The GMRF does not accurately model edges. Choosing  $a_s = 0$  in the pairwise quadratic form gives the pairwise GMRF, and replacing  $|\Delta|^2/2$  by a slower-growing  $\rho(\Delta)$  gives the **pairwise Gibbs distribution** with energy

$$u(x) = \sum_{\{s,r\} \in \mathcal{P}} b_{s,r} \rho(x_s - x_r).$$

- The pairwise Gibbs distribution is the distribution of an MRF, for any  $\rho$ . Not every MRF is a pairwise Gibbs distribution; the general case waits for Chapter 14.
- The most probable pixel given its neighbors solves  $\sum_{r \in \partial s} b_{s,r} \rho'(\hat{x}_s - x_r) = 0$ . So  $\rho$  is potential energy and  $\rho'$  is the force  $x_r$  exerts on  $\hat{x}_s$  — the **influence function**.
- Gaussian  $\rho(\Delta) = |\Delta|^2$ : influence is unbounded, so a pixel across an edge has unbounded effect and edges blur. Influence that falls back toward zero forces **nonconvexity**, and the price is MAP estimates that become discontinuous and unstable.

## 4b. Convex potentials and the scale parameter (Ch. 6.3–6.4)

- A potential is **convex** if  $\rho(\lambda a + (1 - \lambda)b) \leq \lambda \rho(a) + (1 - \lambda)\rho(b)$ . Two advantages: stable MAP estimates, and a convex optimization problem with a tractable global optimum.
- **Total variation**  $\rho = |\Delta|$  is convex and edge preserving, but not strictly convex, and gradient-based algorithms are not convergent on it. **GGMRF**  $\rho = |\Delta|^p/p$  is strictly convex and continuously differentiable for  $p > 1$ . **Huber** and **log cosh** are quadratic near zero and linear far out.
- **QGMRF**:  $|\Delta|^p/p$  at high contrast,  $|\Delta|^q/(pT^{q-p})$  at low contrast, convex for  $1 \leq p < q$ ; with  $q = 2$  it has continuous first and second derivatives.
- **Scale parameter**: use  $\rho((x_s - x_r)/\sigma_x)$  with  $z(\sigma_x) = z_0 \sigma_x^N$ . For the GGMRF the ML estimate is closed form,  $\hat{\sigma}_x^p = \frac{1}{N} \sum b_{s,r} |x_s - x_r|^p$ . Too large means underregularized; too small means overregularized.

## 5. MAP with a non-Gaussian prior (Ch. 7)

- MAP estimation is three steps: pick the forward model, pick the prior model, and minimize

$$f(x; y) = -\log p_{y|x}(y|x) - \log p(x)$$

over the convex set  $\Omega$ , usually  $\mathbb{R}^{+N}$  (positivity).

- Convex  $f$ : every local minimum is global. Strictly convex  $f$ : the minimum is unique and, by the stability theorem, a continuous function of  $y$ . Nonconvex  $f$  lets the estimate jump.
- Optimization is repeated line search.  $x^*$  is a global minimum if and only if it is a fixed point of the line search for every direction  $d$ ; gradient descent, conjugate gradient, and ICD differ only in the choice of  $d$ .
- The line-search function is exactly

$$c + \theta_1 \alpha + \frac{1}{2} \theta_2 \alpha^2 + \sum_{\{s,r\} \in \mathcal{P}} b_{s,r} \rho(\Delta x_{s,r} + \alpha \Delta d_{s,r}),$$

with  $\theta_1 = -e^t \Lambda A d$ ,  $\theta_2 = d^t A^t \Lambda A d$ , and  $e = y - A x$ .

- ICD converges for strictly convex, continuously differentiable  $f$ , and handles positivity by restricting  $\alpha \geq -x_s$ . It stalls on nondifferentiable costs such as total variation.

## 6a. Surrogate functions (Ch. 8.1–8.2)

- A surrogate function  $q(x; x')$  satisfies the two conditions

$$f(x') = q(x'; x'), \quad f(x) \leq q(x; x').$$

- **Majorization minimization:**  $x^{(k+1)} = \arg \min_{x \in \mathbb{R}^N} \{q(x; x^{(k)})\}.$

- The chain

$$f(x^{(k+1)}) \leq q(x^{(k+1)}; x^{(k)}) \leq q(x^{(k)}; x^{(k)}) = f(x^{(k)})$$

gives monotone decrease. The cost values converge if  $f$  is bounded below, but the iterates  $x^{(k)}$  are not guaranteed to converge.

- With the shift, the definition reads  $f(x) \leq Q(x; x') - Q(x'; x') + f(x')$ .
- Properties: monotonicity, gradient equality at  $x'$ , fixed points are stationary points of  $f$ , additivity, and composition.

## 6b. Surrogates for the MAP cost (Ch. 8.3–8.4)

- The data term of the MAP cost is already quadratic; only the prior needs a surrogate. Additivity and composition then give  $Q(x; x')$  from a surrogate potential  $\rho(\Delta; \Delta')$ .
- **Maximum curvature:**  $a_1 = \rho'(\Delta')$ ,  $a_2 = c_{\max}$ . Too conservative when  $\rho''$  is large near zero, and unusable when it is infinite.
- **Symmetric bound:**  $a_2 = \rho'(\Delta')/\Delta'$ , with the limiting value  $\rho''(0)$  at  $\Delta' = 0$ . Valid when the influence function is monotone increasing near zero and concave for  $\Delta > 0$ .
- **Reweighting view:**  $\tilde{b}_{s,r}$  is recomputed each iteration, and  $Q$  is a quadratic cost of the Chapter 5 form.
- **Majorized line search:**  $\alpha^* = \text{clip}\{-\theta_1/\theta_2, [a, b]\}$  in one step, with  $f$  guaranteed to decrease.
- Huber and QGGMRF have bounded  $\rho''(0)$ ; total variation and GGMRF with  $1 \leq p < 2$  do not. Chapter 9 handles those.

## 7a. Duality and the augmented Lagrangian (Ch. 9.1–9.3)

- General problem  $p^* = \inf_{x \in \Omega, Cx=b} f(x)$ ; Lagrangian  $L(x; \lambda) = f(x) + \lambda^t(Cx - b)$ ; dual  $g(\lambda) = \inf_{x \in \Omega} L(x; \lambda)$ , always concave.
- $d^* = \sup_{\lambda} g(\lambda)$  and  $p^* \geq d^*$  by the max-min inequality. Strong duality is  $p^* = d^*$ . Slater's conditions plus convexity give strong duality and the existence of  $\lambda^*$ .
- The dual function  $g(\lambda)$  usually has no explicit form, so  $\lambda^*$  cannot be computed by maximizing it directly. **Augmented Lagrangian:**

$$L(x; a, \lambda) = f(x) + \lambda^t(Cx - b) + \frac{a}{2} \|Cx - b\|^2.$$

- Algorithm:  $x^{(k+1)} = \arg \min_{x \in \Omega} L(x; a, \lambda^{(k)})$  and  $\lambda^{(k+1)} = \lambda^{(k)} + a(Cx^{(k+1)} - b)$ , with  $\lambda^{(k)} \rightarrow \lambda^*$ .
- Convergence does not depend on  $a$ ; large  $a$  means fewer iterations but a more ill-conditioned minimization at each one.

## 7b. Proximal maps, splitting, ADMM (Ch. 9.4–9.5)

- **Proximal map:**  $H(y) = \arg \min_{x \in \Omega} \{\frac{1}{2} \|y - x\|^2 + u(x)\}$ . The cost is strictly convex, so  $H(y)$  is always well defined, and  $H(y)$  is the MAP estimate for denoising  $Y = X + W$  with unit noise variance and  $u(x) = -\log p(x)$ .
- For  $u(x) = \lambda \|x\|_1$  the map separates and equals **shrinkage**,  $S_\lambda(y_i) = \text{sign}(y_i) \max\{|y_i| - \lambda, 0\}$ .
- **Variable splitting:** for  $\hat{x} = \arg \min_{x \in \Omega} \{f(x) + g(Bx)\}$ , set  $v = Bx$  and minimize  $f(x) + g(v) + \frac{a}{2} \|Bx - v + u\|^2$  jointly, then update  $u \leftarrow u + B\hat{x} - \hat{v}$ . A single iteration of the alternating minimization gives **ADMM**, which still has guaranteed convergence.
- **Total variation:** split  $v = Bx$  in  $\frac{1}{2} \|y - Ax\|_\Lambda^2 + \|Bx\|_1$ . The  $x$  update is quadratic; the  $v$  update is the  $L_1$  proximal map, so  $\hat{v} \leftarrow S_{1/a}(B\hat{x} + u)$ .
- For convex constraints the  $v$  update is  $\hat{v} \leftarrow \text{clip}\{\hat{x} + u, \mathcal{A}\}$ . Splitting pays off when  $f$  and  $g$  need different algorithms.

## 8a. Plug-and-play (Ch. 10.1–10.2)

- Regularized inversion is  $x^* = \arg \min_x \{f(x) + h(x)\}$ ; MAP is  $f = -\log p(y|x)$ ,  $h = -\log p(x)$ . It is not modular, and it is objective-based.
- The proximal maps  $F$  and  $H$  convert objectives into actions;  $\sigma^2$  is a selectable parameter.
- **PnP-ADMM**:

$$x \leftarrow F(v - u), \quad v \leftarrow H(x + u), \quad u \leftarrow u + (x - v).$$

With proximal maps this is exactly ADMM; with general actions it is no longer solving an optimization problem.

- $H(x)$  has the form of a MAP denoiser for additive white Gaussian noise of variance  $\sigma^2$ , so it can be trained from synthetic pairs  $(V_n, X_n)$  using  $L(\theta) = \sum_n \|V_n - H_\theta(X_n)\|^2$ .
- **Moreau**:  $H$  is a proximal map if and only if it is nonexpansive and conservative; a conservative  $H$  must also have a self-adjoint gradient,  $\nabla H(x) = [\nabla H(x)]^t$ . A DNN trained for optimal image denoising is not likely to be a proximal map, but it could serve as an excellent prior model in a PnP framework.

## 8b. Consensus equilibrium and MACE (Ch. 10.3–10.4)

- Setting  $x = v$  at the convergence point of PnP-ADMM gives the **CE equations**

$$x^* = F(x^* - u^*), \quad x^* = H(x^* + u^*),$$

which replace optimization as the defining criterion for the solution.

- If  $F$  and  $H$  are proximal maps of proper closed convex  $f$  and  $h$ , then  $x^*$  solves the CE equations if and only if  $x^*$  is a global minimum of  $f(x) + h(x)$ . For general agents the CE equations are a force balance equilibrium that does not result from minimizing any underlying cost function.
- Serial solution:  $w_1^*$  is the CE solution if and only if it is a fixed point of  $T = (2H - I)(2F - I)$ . Douglas–Rachford with Mann iterations converges when  $T$  is nonexpansive; at  $\rho = 0.5$  this is exactly PnP-ADMM.
- MACE**:  $K$  heterogeneous agents with weights  $\mu_k$ ,  $1 = \sum_{k=1}^K \mu_k$ , satisfy

$$\mathbf{F}(\mathbf{w}^*) = \mathbf{G}_\mu(\mathbf{w}^*), \quad x^* = \sum_{k=1}^K \mu_k w_k^*.$$

Since  $\mathbf{G}_\mu \mathbf{G}_\mu = \mathbf{G}_\mu$ , the solution is a fixed point of  $\mathbf{T} = (2\mathbf{G} - I)(2\mathbf{F} - I)$ .

## 9a. The EM algorithm (Ch. 12.1–12.4)

- Model parameters must be estimated from indirect or incomplete measurements, because the undistorted image is not available. ML on the incomplete data is consistent and asymptotically efficient, but  $p(y|\theta) = \sum_x p(y|x, \theta)p(x|\theta)$  has no simple closed form.
- Missing data  $X$ ; complete data  $(X, Y)$ ; incomplete data  $Y$ . **Decomposition:**

$$\log p(y|\theta) = Q(\theta; \theta') + H(\theta; \theta'),$$

with  $Q$  the conditional expectation of  $\log p(y, X|\theta)$  and  $H$  the negative conditional expectation of  $\log p(X|y, \theta)$ .

- Jensen's inequality gives  $H(\theta; \theta') \geq H(\theta'; \theta')$ , hence the key bound

$$\log p(y|\theta) - \log p(y|\theta') \geq Q(\theta; \theta') - Q(\theta'; \theta'),$$

so  $Q$  is a **surrogate function** for maximizing  $\log p(y|\theta)$ .

- EM:** the E-step computes  $Q(\theta; \theta^{(k)})$ ; the M-step sets  $\theta^{(k+1)} = \arg \max_{\theta \in \Omega} Q(\theta; \theta^{(k)})$ . The log likelihood is monotone increasing.

## 9b. EM in practice, and the exponential-family rule (Ch. 12.5–12.7)

- Gaussian mixture,  $M$  classes selected by a switch. The E-step assigns  $P_{n,m}$  from Bayes' rule; the M-step assigns  $N_m \leftarrow \sum_n P_{n,m}$ ,  $\pi_m \leftarrow N_m/N$ ,  $\mu_m \leftarrow \frac{1}{N_m} \sum_n y_n P_{n,m}$ , and  $\sigma_m^2 \leftarrow \frac{1}{N_m} \sum_n (y_n - \mu_m)^2 P_{n,m}$ .
- The assignment is soft. EM generalizes the hard classification heuristic: E-step = soft classification, M-step = ML estimation given those soft classifications. Iterations can still be trapped in a local maximum.
- A  $k$ -parameter **exponential family** has  $p(y|\theta) = \exp\{\langle \eta(\theta), T(y) \rangle + d(\theta) + s(y)\}$ , with  $T(y)$  the natural sufficient statistic, and then  $\hat{\theta}_{ML} = f(T(y))$ .
- If  $p(y, x|\theta)$  is an exponential family, then  $Q(\theta; \theta') = \langle g(\theta), \bar{T}(y) \rangle + d(\theta) + \text{const.}$  with  $\bar{T}(y) = \mathbb{E}[T(y, X) | Y = y, \theta']$ , giving the **EM update**

$$\theta'' = f(\bar{T}(y)),$$

the ML estimate with  $\bar{T}$  in place of  $T$ .

## 10a. Markov chains and HMMs (Ch. 13.1–13.3)

- A Markov chain satisfies  $P\{X_n = j | X_k \text{ for all } k < n\} = P\{X_n = j | X_{n-1}\}$  on a countable state space  $\Omega = \{0, \dots, M-1\}$ , and

$$p(x) = \tau_{x_0} \prod_{n=1}^N P_{x_{n-1}, x_n}.$$

- Natural sufficient statistics  $N_j$  and  $K_{i,j}$  give the ML estimates  $\hat{\tau}_j = N_j$  and  $\hat{P}_{i,j} = K_{i,j} / \sum_j K_{i,j}$  — count transitions and normalize each row.
- An **HMM** is doubly stochastic: the unobserved chain  $X_n$  controls the observed process  $Y_n$ , with

$$p(y, x | \theta) = \tau_{x_0} \prod_{n=1}^N \{f(y_n | x_n) P_{x_{n-1}, x_n}\}.$$

- MAP state estimation looks like a search over  $M^{N+1}$  sequences, but **dynamic programming** reduces it to order  $NM^2$ . The **forward-backward** algorithm computes  $\alpha_n(j)$  and  $\beta_n(j)$ , also in order  $NM^2$ .
- EM training replaces the natural sufficient statistics with their conditional expectations; the posteriors needed in the E-step are exactly what forward-backward computes.

## 10b. Ergodicity and detailed balance (Ch. 13.4–13.5)

- The marginal density evolves as  $\pi^{(n+1)} = \pi^{(n)}P$ , so  $\pi^{(n)} = \pi^{(0)}P^n$ . If a steady state exists and does not depend on  $\pi^{(0)}$ , then  $\pi^\infty = \pi^\infty P$ .
- **Limit Theorem:**  $\Omega$  finite, irreducible, and aperiodic imply a unique stationary distribution  $\pi_j = \lim_{n \rightarrow \infty} [P^n]_{i,j} > 0$ , the unique solution of the full-balance equations. Such chains are called **ergodic**, and for them expectations can be replaced by time averages.
- Time reversal: if  $Y_n = X_{N-n}$  then  $Y_n$  is a Markov chain, and in steady state  $\pi_i P_{i,j} = \pi_j Q_{j,i}$ .
- **Reversible** means  $P_{i,j} = Q_{i,j}$ , which gives the detailed-balance equations

$$\pi_i P_{i,j} = \pi_j P_{j,i} :$$

the rate of transitions from  $i$  to  $j$  equals the rate from  $j$  to  $i$ .

- Summing detailed balance over  $i$  gives full balance, so a detailed-balance solution is the stationary distribution of an ergodic Markov chain.

# 11. General MRFs (Ch. 14)

- A neighborhood system requires  $r \in \partial s \Rightarrow s \in \partial r$  and  $s \notin \partial s$ .  $X$  is an MRF if  $P\{X_s = x_s | X_r, r \neq s\} = P\{X_s = x_s | X_{\partial s}\}$ . The joint distribution is *not* the product of the conditionals.
- A clique is a set of points that are all neighbors of one another, and a Gibbs distribution is  $p(x) = \frac{1}{z} \exp\{-\sum_{c \in \mathcal{C}} V_c(x_c)\}$  with energy function  $u(x) = \sum_{c \in \mathcal{C}} V_c(x_c)$ .
- **Hammersley–Clifford theorem:** let  $S$  be an  $N$ -point lattice with neighborhood system  $\partial s$ , and let  $X$  be a random process on  $S$  with  $p(x) > 0$ . Then  $X$  is an MRF if and only if  $p(x)$  is a Gibbs distribution.
- $z$  is a sum over exponentially many states, so ML parameter estimation is blocked; but  $\hat{x} = \arg \min_x \{u(x)\}$  and  $p(x')/p(x) = \exp\{u(x) - u(x')\}$  need only  $u$ .
- **Ising model:**

$$p(x) = \frac{1}{z(\beta)} \exp \left\{ -\beta \sum_{\{r,s\} \in \mathcal{C}} \delta(x_r \neq x_s) \right\}, \quad \beta = \frac{2J}{k_b T}.$$

In 2-D, for  $\beta > \beta_c \approx 0.88$ , correlations extend to infinite distances.

## 12. Stochastic simulation (Ch. 15)

- The objective is to generate a sequence  $X^{(k)}$  whose distribution converges to the desired Gibbs distribution  $p(x)$ . The **Metropolis** sampler draws a proposal  $W$  from  $q(w|X^{(k)})$ , which must obey the symmetry property  $q(w|x) = q(x|w)$ , and accepts it with probability

$$\alpha(w|x^{(k)}) = \min \left\{ 1, e^{-[u(w) - u(x^{(k)})]} \right\}.$$

- The sequence  $\{X^{(k)}\}$  is a Markov chain, and the acceptance rule makes it satisfy the detailed-balance equations, so its stationary distribution is  $p(x) = \frac{1}{Z} e^{-u(x)}$ .
- The **Hastings–Metropolis** sampler allows proposal distributions that are not symmetric, correcting the bias through the acceptance probability

$$\alpha(w|x^{(k)}) = \min \left\{ 1, \frac{q(x^{(k)}|w)}{q(w|x^{(k)})} \frac{p(w)}{p(x^{(k)})} \right\}.$$

- The **Gibbs sampler** is the coordinatewise Hastings–Metropolis sampler whose proposal for pixel  $s$  is the conditional distribution of  $X_s$  given its neighbors,  $X_s^{(k+1)} \sim p_s(x_s|X_{\partial s}^{(k)})$ . Then  $\alpha = \min\{1, 1\} = 1$ : all proposals are accepted.
- Convergence is guaranteed, but it can be extremely slow.

# Summary

- **One core idea:** model the sensor, model the image, and compute the reconstruction, most often as  $\hat{x}_{MAP} = \arg \min_{x \in \Omega} \{-\log p_{y|x}(y|x) - \log p_x(x)\}$ .
- **The models grew** from causal to noncausal, and from Gaussian to non-Gaussian to discrete:  $p(x) \propto \exp\{-\frac{1}{2}x^t A^t \Lambda^{-1} A x\}$ , then  $\exp\{-\frac{1}{2}x^t B x\}$ , then  $\frac{1}{z} \exp\{-u(x)\}$ .
- **The optimization tools grew** from the closed-form solution  $\mu(y) = (A^t A + \sigma^2 B)^{-1} A^t y$ , to surrogate functions with  $f(x) \leq q(x; x')$ , to variable splitting and ADMM, to the equilibrium equations  $x^* = F(x^* - u^*)$  and  $x^* = H(x^* + u^*)$ .
- **EM handles incomplete data** through the decomposition  $\log p(y|\theta) = Q(\theta; \theta') + H(\theta; \theta')$ , in which  $Q$  is a surrogate function for the log likelihood.
- **Markov chain theory serves two purposes:** it models data through  $p(x) = \tau_{x_0} \prod_n P_{x_{n-1}, x_n}$ , and it drives the samplers through detailed balance  $\pi_i P_{i,j} = \pi_j P_{j,i}$ .

# Final exam

- The final exam is during finals week, Dec. 14–18. The time and location will be announced by the Registrar.
- Exams are closed book.

|                               |     |
|-------------------------------|-----|
| Homework and laboratories     | 30% |
| Midterm                       | 30% |
| Final exam (may be a project) | 40% |

Thank you for a great semester.