Chapter V
Image Restoration

Images are often degraded during the data acquisition process. The degradation may involve blurring, information loss due to sampling, quantization effects, and various sources of noise. The purpose of image restoration is to estimate the original image from the degraded data. Applications range from medical imaging, astronomical imaging, to forensic science, etc. Often the benefits of improving image quality to the maximum possible extent far outweigh the cost and complexity of the restoration algorithms involved.

1 Degradation Model

The most general degradation model is that of a conditional pdf for the data $\mathbf{y}$ given the original image $\mathbf{x}$, as depicted in Fig. 1. The domains of $\mathbf{x}$ and $\mathbf{y}$ are generally (but not always) discrete. For instance, $\mathbf{x}$ and $\mathbf{y}$ could be images with the same number N of pixels.

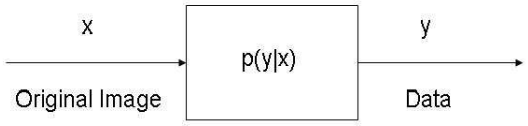


Figure 1: General statistical model for image restoration.

We consider five models that are representative of actual image restoration problems or at least are useful mathematical abstractions thereof.

Model 1 (see Fig. 2): Additive white noise

$$\mathbf{y} = \mathbf{x} + \mathbf{w}$$

Model 2 (see Fig. 3): Linear blur plus additive white noise

$$\mathbf{y} = H\mathbf{x} + \mathbf{w}$$

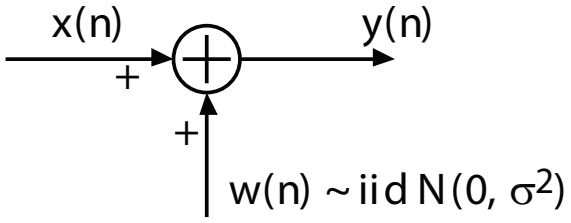


Figure 2: Additive White Gaussian Noise (AWGN) degradation model.

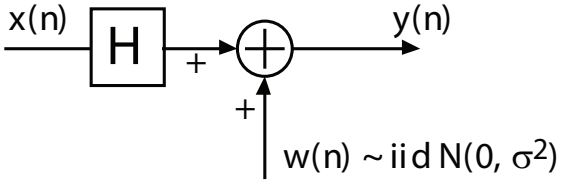


Figure 3: Blur + AWGN degradation model.

where H represents the effects of camera or object motion, atmospheric turbulence, optics, etc.

Model 3: Tomography [1, Ch. 10] [2, p. 115] [3].

Consider the following imaging system for transmission tomography [1, Ch. 10]. An object (typically a slice of a patient's body) is irradiated along direction θ by an X-ray or gamma-ray source. These high-energy photons travel through the object and are subsequently detected and counted. At each location (x, y) inside the object, the photon is subject to possible capture, with probability $f(x, y)dl$ over an elementary path segment of length dl. The intensity of the surviving photons that travelled along flight path L is therefore given by

$$\lambda(L) = \lambda_0 \exp\left\{-\int_L f(x, y)\, dl\right\}. \tag{1}$$

where λ_0 is the source intensity in the direction of L.

For a flight path L with coordinates (s, θ) (see Fig. 5), the normalized log-intensity at the detector is given by

$$g(s, \theta) = -\ln \frac{\lambda(L)}{\lambda_0} = \int_L f(x, y)\, dl.$$

A first problem is to recover the absorption map $f(x, y)$ from these line integrals. We write $g = \mathscr{R} f$ where $\mathscr{R}$ is the *Radon transform*, defined as

$$g(s, \theta) = \int\int_{\mathbb{R}^2} f(x, y)\delta(-x \sin\theta + y \cos\theta)\, dxdy, \quad s \in \mathbb{R},\ 0 \le \theta < \pi,$$

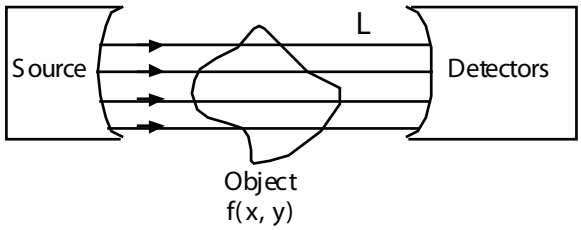


Figure 4: Irradiation of body in transmission tomography.

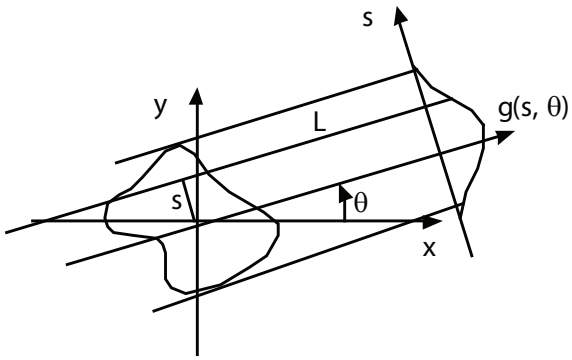


Figure 5: Parameterization of flight path L by (s, θ).

and $\delta(\cdot)$ is the Dirac impulse.

For high-count imaging modalities, the quantum nature of the measurements is often ignored, and the problem becomes one of inverting the Radon transform. In low-count imaging modalities, a more accurate, physics-based model taking into account Poisson statistics is desirable. The data collected by the detectors are independent Poisson random variables

$$y(s, \theta) \sim \text{Poisson}[\lambda(s, \theta)]$$

where $\lambda(s, \theta)$ is given in (1).

Similar models apply to other imaging modalities such as PET (positron-electron tomography), where positron-electron annihilation produces two high-energy photons traveling in opposite directions, see Fig. 6.

Another example from photon-limited imaging involving a CCD array is shwon in Fig. 7.

Model 4: Inverse problem.

Here

$$\mathbf{y} = H\mathbf{x}$$

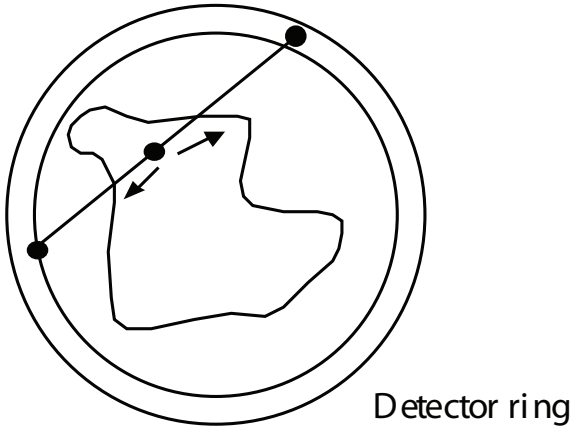


Figure 6: PET imaging.

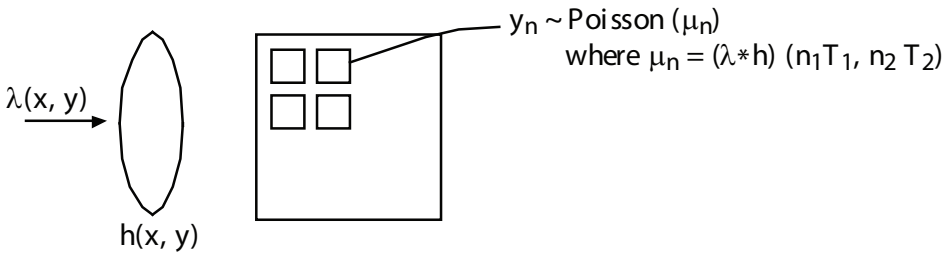


Figure 7: CCD imaging.

where H is a many-to-one mapping. Of course this may be viewed as a special case of Fig. 3 in which the noise variance is zero. The observational model $p(\mathbf{y}|\mathbf{x})$ takes the form of a Dirac impulse:

$$p(\mathbf{y}|\mathbf{x}) = \delta(\mathbf{y} - H\mathbf{x}).$$

The fact that H is a many-to-one mapping creates difficulties for estimating $\mathbf{x}$ even in the absence of noise, due to the nonuniqueness of the solution. Among the many possible solutions, we will seek the one that optimizes an appropriate criterion. This problem is also called *image recovery*.

2 Inverse Problem

Consider Model 4, $\mathbf{y} = H\mathbf{x}$, where H is linear and invertible. In the absence of noise, the inverse filter solution

$$\hat{\mathbf{x}} = H^{-1}\mathbf{y}$$

does perfectly recover the original image $\mathbf{x}$. To study the behavior of $\hat{\mathbf{x}}$ in the presence of noise (Model 2), first observe that the reconstruction error is given by

$$\mathbf{e} = \hat{\mathbf{x}} - \mathbf{x} = H^{-1}\mathbf{w}.$$

The case of linear shift-invariant H ilustrates the problem with $\hat{\mathbf{x}}$. If $H(f)$ is very small at some frequencies (as is the case with common blur functions), $\hat{\mathbf{x}}$ has the unfortunate property that even a small amount of noise will undergo dramatic amplification through the inverse filter. For Gaussian blur for instance, $H^{-1}(f)$ increases exponentially fast with f, see Fig. 8.

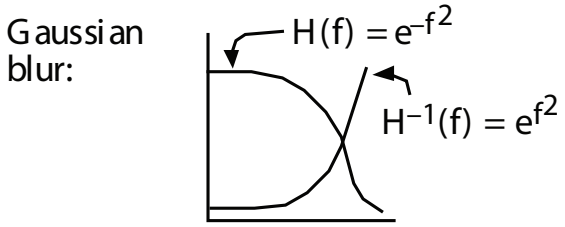


Figure 8: A blur filter $H(f)$ and the corresponding inverse filter.

If $w(n)$ is white noise with variance σ^2, then the spectral density of the reconstruction error, $S_e(f) = \sigma^2|H^{-1}(f)|^2$, is unacceptably large at high frequencies.

The same issue appears in several variations of the problem. If H is linear but many-to-one (e.g., samples of the Radon transform are observed), there exists an affine subspace $\mathcal{S}(\mathbf{y})$ such that $H\mathbf{x} = \mathbf{y}$ for all $\mathbf{x} \in \mathcal{S}(\mathbf{y})$. The minimum-norm solution is given by $\hat{\mathbf{x}} = H^{\dagger}\mathbf{y}$, where is the so-called *pseudo-inverse filter*, a projection operator.

Similar problems arise if H is nonlinear.

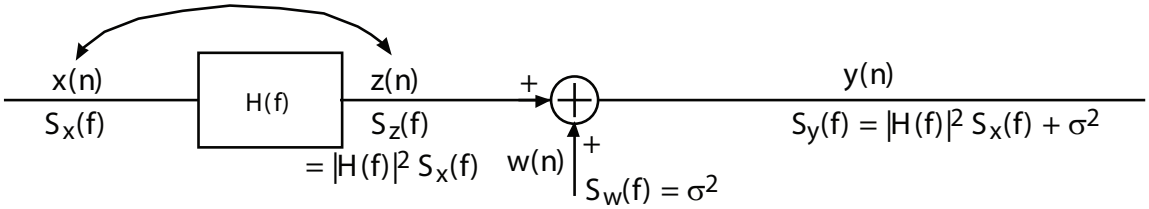
3 Wiener Filter

This section studies a well-known linear estimator, the Wiener filter, which minimizes the mean-squared error (MSE) of the estimated image for linear Gaussian models. Specifically, consider Model 2, with a LSI system $H(f)$ and stationary random processes X and W. The processes X and W are assumed to be uncorrelated.

The spectral density of Y and the cross-spectral density of X and Y are respectively given by

$$S_Y(f) = |H(f)|^2 S_X(f) + \sigma_W^2 \quad (2)$$
$$S_{XY}(f) = H^*(f) S_X(f) \quad (3)$$

see Fig. 3.



Design a finite-impulse response (FIR) filter with coefficients $\mathbf{g} = \{g(m),\, m \in \mathcal{M}\}$ and let

$$\hat{X}(n) = \sum_k g(k) Y(n-k), \quad \forall n \in \mathbb{Z}^2.$$

The Wiener filter is $\mathbf{g}$ that minimizes the expected mean-squared reconstruction error,

$$\mathcal{E}(\mathbf{g}) = \mathbb{E}[|\hat{X}(n) - X(n)|^2] = \mathbb{E}\left[\left|\sum_k g(k) Y(n-k) - X(n)\right|^2\right], \quad \forall n \in \mathbb{Z}^2.$$

The optimal filter coefficients are given by

$$\begin{aligned} 0 &= \frac{\partial \mathcal{E}(\mathbf{g})}{\partial g(m)} \\ &= \mathbb{E}\left[\left(\sum_{k\in\mathcal{M}} g(k) Y(n-k) - X(n)\right) Y(n-m)\right] \\ &= \sum_{k\in\mathcal{M}} g(k) R_Y(m-k) - R_{XY}(m), \quad m \in \mathcal{M}. \end{aligned} \tag{4}$$

This is a linear system with $|\mathcal{M}|$ equations and $|\mathcal{M}|$ unknowns $\{g(k),\, k \in \mathcal{M}\}$. The solution is

$$\mathbf{g}_{opt} = \mathsf{R}_Y^{-1} \mathbf{r}_{XY}$$

where R_Y is the symmetric Toeplitz matrix with element (m,k) given by $R_Y(m-k)$, and $\mathbf{r}_{XY}$ is the vector with element m given by $R_{XY}(m)$. The resulting MSE is given by

$$\mathcal{E}_{\text{opt}} = \mathcal{E}(\mathbf{g}_{opt}) = \sigma_X^2 - \mathbf{r}_{XY}^T \mathsf{R}_Y^{-1} \mathbf{r}_{XY}.$$

If $G(f)$ is an IIR filter ($\mathcal{M} = \mathbb{Z}^2$), Wiener filtering admits a convenient frequency interpretation. Taking the DSFT of (4), we obtain

$$G(f) S_Y(f) = S_{XY}(f)$$

hence, from (2) and (3),

$$G(f) = \frac{H^*(f) S_X(f)}{|H(f)|^2 S_X(f) + \sigma^2}.$$

For the linear degradation model problem considered above, the Wiener filter is in fact the estimator that minimizes the reconstruction MSE among *all* possible estimators (linear and nonlinear), when the processes X and W are Gaussian. However, when those conditions are not met, the Wiener solution presents the following drawbacks.

1. linear estimation is a restriction; in general, a nonlinear estimator would outperform the Wiener filter.
2. the framework applies to linear degradation models only.
3. restored images are typically oversmoothed due to the use of the MSE criterion, which penalizes all components of the reconstruction error equally, while the HVS has higher tolerance for high-frequency errors.
4. image statistics are often unknown.

For more general image restoration problems, we seek nonlinear estimators that apply to arbitrary models and enjoy optimality properties beyond MSE. We also seek a systematic way to deal with unknown statistics.

4 Maximum-Likelihood Estimation

The Maximum-Likelihood (ML) estimation method is applicable to arbitrary degradation model $p(\mathbf{y}|\mathbf{x})$ [4]. Assume that $\mathbf{y}$ is a N-vector (e.g., N-pixel image) and that $\mathbf{x}$ is a deterministic but unknown image that is completely described by some K-dimensional parameter θ. Then estimating $\mathbf{x}$ from the data $\mathbf{y}$ is equivalent to estimating θ from $\mathbf{y}$.

Examples:

1. Constant image: $x(n) = \theta$, for all n $(K = 1)$
2. Planar patch: $x(n) = \theta_1 + \theta_2 n_1 + \theta_3 n_2$ $(K = 3)$

4.1 Fisher Information

Before studying the properties of ML estimators, we state the celebrated Cramer-Rao lower bound on the variance of any unbiased estimator of θ given data $\mathbf{y}$ [4].

For $K = 1$: For any unbiased estimator,

$$\mathrm{Var}(\hat{\theta}) = \mathbb{E}(\hat{\theta}(\mathbf{Y}) - \theta)^2 \geq I_\theta^{-1}$$

where, under some regularity conditions on the model $p(\mathbf{Y}|\theta)$,

$$I_\theta = \mathbb{E}\left(\frac{\partial \ln p(\mathbf{Y}|\theta)}{\partial \theta}\right)^2$$

is the *Fisher information* for θ. Under additional regularity conditions on $p(\mathbf{Y}|\theta)$, we have

$$I_\theta = \mathbb{E}\left(-\frac{\partial^2 \ln p(\mathbf{Y}|\theta)}{\partial \theta^2}\right). \tag{5}$$

For $K \geq 1$:

$$\mathrm{Cov}(\hat{\theta}) = \mathbb{E}[(\hat{\theta}-\theta)(\hat{\theta}-\theta)^T] \geq \mathsf{I}_\theta^{-1} \tag{6}$$

where

$$\begin{aligned} \mathsf{I}_\theta &= \mathbb{E}[(\nabla_\theta \ln p(\mathbf{Y}|\theta))(\nabla_\theta \ln p(\mathbf{Y}|\theta))^T] \\ &= \mathbb{E}[\nabla_\theta^2 \ln p(\mathbf{Y}|\theta)] \end{aligned}$$

is (under regularity conditions) the Fisher information matrix for θ.

The matrix inequality notation $\mathsf{A} \geq \mathsf{B}$ means that $\mathsf{A}-\mathsf{B}$ is nonnegative definite.

An estimator which attains the Cramer-Rao bound (6) is said to be efficient.

4.2 Desirable Properties of an Estimator

Some desirable properties of any estimator $\hat{\theta}(\mathbf{y})$ of θ are the following. We shall soon examine which ones of these properties apply to ML estimators.

(a) Unbiasedness: $\mathbb{E}[\hat{\theta}(\mathbf{Y})] = \theta$.

(b) Consistency in probability: $\hat{\theta}(\mathbf{Y}) \to \theta$ in probability, as $N \to \infty$

(c) m.s. consistency: $\hat{\theta}(\mathbf{Y}) \to \theta$ m.s. as $N \to \infty$

(d) Efficiency: $\mathrm{Cov}(\hat{\theta}) = \mathsf{I}_\theta^{-1}$.

4.3 Example #1

Our first example consists of a constant image under the AWGN model:

$$Y(n) \sim \text{iid } \mathcal{N}(\theta, \sigma^2), \quad 1 \leq n \leq N.$$

Hence

$$\begin{aligned} p(\mathbf{y}|\theta) &= \prod_{n=1}^{N}\left((2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{(y(n)-\theta)^2}{2\sigma^2}\right\}\right) \\ \ln p(\mathbf{y}|\theta) &= -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y(n)-\theta)^2 \end{aligned}$$

which is quadratic in θ. We have

$$\frac{d \ln p(\mathbf{y}|\theta)}{d\theta} = -\frac{1}{\sigma^2} \sum_{n=1}^{N} (\theta - y(n)) = -\frac{1}{\sigma^2} \left(N\theta - \sum_{n=1}^{N} y(n) \right) \tag{7}$$

The solution to the likelihood equation $0 = \frac{d \ln p(\mathbf{y}|\theta)}{d\theta}$ is the ML estimator:

$$\hat{\theta}(\mathbf{y}) = \frac{1}{N} \sum_{n} y(n).$$

which is simply the mean of the observations. Hence

$$\mathbb{E}[\hat{\theta}(\mathbf{Y})] = \frac{1}{N} \sum_{n} \mathbb{E}[Y(n)] = \theta$$

($\hat{\theta}$ is unbiased). Moreover

$$\mathbb{E}(\hat{\theta}(\mathbf{Y}) - \theta)^2 = \mathbb{E}\left[\frac{1}{N} \sum_{n} (Y(n) - \theta)^2 \right] = \frac{\sigma^2}{N}$$

which tends to zero as $N \to \infty$; hence $\hat{\theta}$ is m.s. consistent.

Fisher information: differentiating (7) with respect to θ, we have

$$\frac{d^2 \ln p(\mathbf{y}|\theta)}{d\theta^2} = -\frac{N}{\sigma^2}.$$

Applying (5), we obtain

$$I_\theta = \frac{N}{\sigma^2}.$$

It is intuitively pleasing that this "information" coincides with the signal-to-noise ratio N/σ^2. Moreover, the Cramer-Rao bound is satisfied with equality here, so $\hat{\theta}$ is efficient.

4.4 Properties of ML Estimators

Under some technical conditions on $p(\mathbf{y}|\theta)$, ML estimates satisfy the following properties [4]:

(a) asymptotically unbiased: $\lim_{N\to\infty} \mathbb{E}[\hat{\theta}(\mathbf{Y})] = \theta$.

(b) consistent in probability: $\hat{\theta}(\mathbf{Y}) \to \theta$ in probability.

(c) asymptotically efficient: $\lim_{N\to\infty} \mathrm{Cov}(\hat{\theta}) = \mathsf{I}_\theta^{-1}$.

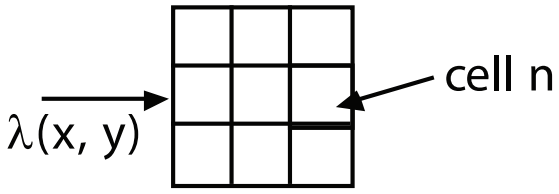
Despite its attractive asymptotic properties, ML may not be the best estimator for finite N. In fact, there is no guarantee that ML is good at all for small N.

4.5 Example #2

Computing ML estimates is often difficult, even in seemingly simple problems. Consider the following example. A CCD array is illuminated by light with spatially-varying intensity,

$$\lambda(t) = \theta_1\phi_1(t) + \theta_1\phi_1(t), \quad t \in \mathbb{R}^2$$

where $\phi_1(t)$ and $\phi_2(t)$ are known intensity functions and θ_1 and θ_2 are unknown, positive scaling factors. For instance, ϕ_1 could represent a point source subjected to Gaussian blur; and ϕ_2 could represent a constant background intensity.



The photon count in cell n is $Y_n \sim \text{Poisson}(\lambda_n)$ where

$$\lambda_n = \int_{cell\ n} \lambda(t)\, dt = \theta_1\phi_{1n} + \theta_2\phi_{2n}$$

and

$$\phi_{in} = \int_{cell\ n} \phi_i(t)\, dt, \quad i = 1, 2.$$

The likelihood for θ_1, θ_2 is given by

$$L(\theta) = P(\mathbf{y}|\theta) = \prod_n \left(\frac{\lambda_n^{y_n}}{y_n!}\, e^{-\lambda_n} \right)$$

and the log likelihood by

$$l(\theta) = \ln L(\theta) = \sum_n [y_n \ln \lambda_n - \lambda_n - \ln y_n!]$$

The ML estimate $\hat{\theta}$ satisfies the equations

$$0 = \frac{\partial l(\theta)}{\partial \theta_i} = \sum_n \left[\frac{y_n}{\theta_1\phi_{1n} + \theta_2\phi_{2n}} - 1 \right] \phi_{in}, \quad i = 1, 2. \tag{8}$$

This is a nonlinear system of equations which cannot be solved in closed form.

Several numerical techniques can be considered to solve nonlinear systems of the above form. As an alternative to standard methods of steepest descent, conjugate gradients, etc., we now introduce a statistical optimization algorithm whose effectiveness has been demonstrated in a variety of applications.

5 Expectation-Maximization (EM) Algorithm

This is an iterative algorithm published by Dempster, Laird and Rubin in 1977 [5]. The EM algorithm requires the definition of the following ingredients:

- an incomplete-data space $\mathcal{Y}$ (measurement space)
- a complete-data space $\mathcal{Z}$ (many choices are possible)
- a many-to-one mapping $h \ : \ \mathcal{Z} \rightarrow \mathcal{Y}$.

The incomplete-data and complete-data loglikelihood functions are respectively given by

$$l_{id}(\theta) \triangleq \ln p(\mathbf{Y}|\theta), \quad l_{cd}(\theta) \triangleq \ln p(\mathbf{Z}|\theta).$$

Assume an initial estimate $\hat{\theta}^{(0)}$ is given. The EM algorithm alternates between Expectation (E) and Maximization (M) steps for updating the estimate $\hat{\theta}^{(k)}$ of the unknown parameter θ at iteration $k > 0$.

E-Step: Compute

$$Q(\theta|\hat{\theta}^{(k)}) \triangleq \mathbb{E}[l_{cd}(\theta)|\mathbf{Y}, \hat{\theta}^{(k)}]$$

where the conditional expectation is evaluated assuming that the true θ is equal to $\hat{\theta}^{(k)}$.

M-step:

$$\hat{\theta}^{(k+1)} = \arg\max_{\theta} Q(\theta|\hat{\theta}^{(k)}).$$

To illustrate how the EM algorithm works, appy it to our Poisson imaging example and define complete data $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)$ where $\mathbf{z}_i = \{z_{in}\}$ are the photon counts due to processes 1 and 2, respectively. Thus z_{in} are independent Poisson random variables with intensities $\theta_i \phi_{in}$. The mapping h is given by

$$\mathbf{y} = h(\mathbf{z}) = \mathbf{z}_1 + \mathbf{z}_2.$$

We have

$$l_{cd}(\theta) = \ln p(\mathbf{z}|\theta) = \sum_n [z_{1n} \ln(\theta_1 \phi_{1n}) - \theta_1 \phi_{1n} - \ln z_{1n}! + z_{2n} \ln(\theta_2 \phi_{2n}) - \theta_2 \phi_{2n} - \ln z_{2n}!]$$

E-step:

$$\begin{aligned} Q(\theta|\hat{\theta}^{(k)}) \ &= \ \sum_n \Big\{ \mathbb{E}[Z_{1n}|\mathbf{Y} = \mathbf{y}, \theta^{(k)}] \ \ln(\theta_1 \phi_{1n}) - \theta_1 \phi_{1n} \\ &\qquad + \mathbb{E}[Z_{2n}|\mathbf{Y} = \mathbf{y}, \theta^{(k)}] \ \ln(\theta_2 \phi_{2n}) - \theta_2 \phi_{2n} - \mathbb{E}[\ln Z_{1n}! + \ln Z_{2n}!|\mathbf{Y}, \theta^{(k)}] \Big\} \end{aligned}$$

Defining the ratio

$$r_{in} \triangleq \frac{\theta_i^{(k)} \phi_{in}}{\theta_1^{(k)} \phi_{1n} + \theta_2^{(k)} \phi_{2n}}, \quad i = 1, 2,$$

we obtain

$$\mathbb{E}[Z_{in}|\mathbf{Y}=\mathbf{y},\theta^{(k)}] = \mathbb{E}[Z_{in}|Y_n=y_n,\theta^{(k)}] = \frac{\theta_i^{(k)}\phi_{in}}{\theta_1^{(k)}\phi_{1n}+\theta_2^{(k)}\phi_{2n}}y_n = r_{in}y_n, \quad i=1,2.$$

Hence

$$Q(\theta|\hat{\theta}^{(k)}) = \sum_n [r_{1n}y_n \ln(\theta_1\phi_{1n}) - \theta_1\phi_{1n} + r_{2n}y_n \ln(\theta_2\phi_{2n}) - \theta_2\phi_{2n}] + C$$

where C is a constant, independent of θ.

M-Step: Setting the partial derivatives of $Q(\theta|\hat{\theta}^{(k)})$ with respect to θ_1 and θ_2 to zero, we obtain

$$0 = \frac{Q(\theta|\hat{\theta}^{(k)})}{\partial\theta_i} = \sum_n \left[\frac{r_{in}y_n}{\theta_i} - \phi_{in}\right], \quad i=1,2.$$

Hence

$$\begin{aligned}\hat{\theta}_i^{(k+1)} &= \sum_n r_{in}y_n \Big/ \sum_n \phi_{in} \\ &= \sum_n \frac{\theta_i^{(k)}\phi_{in}}{\theta_1^{(k)}\phi_{1n}+\theta_2^{(k)}\phi_{2n}}y_n \Big/ \sum_n \phi_{in} \quad i=1,2.\end{aligned}$$

Let's see what the sequence $\hat{\theta}_i^{(k)}$ might converge to. If $\hat{\theta}_i^{(k)} \to \hat{\theta}_i^{(\infty)}$ as $k\to\infty$, then

$$\begin{aligned}\hat{\theta}_i^{(\infty)}\sum_n \phi_{in} &= \sum_n \frac{\hat{\theta}_i^{(\infty)}\phi_{in}}{\hat{\theta}_1^{(\infty)}\phi_{1n}+\hat{\theta}_2^{(\infty)}\phi_{2n}}y_n \\ 0 &= \sum_n \left[\frac{y_n}{\hat{\theta}_1^{(\infty)}\phi_{1n}+\hat{\theta}_2^{(\infty)}\phi_{2n}} - 1\right]\phi_{in} \quad i=1,2.\end{aligned}$$

which coincides with the system (8) whose solution yields the true ML estimates.

Properties of the EM algorithm [6]

1. The sequence of incomplete-data likelihoods $l_{id}^{(k)}$ is nondecreasing with iteration number k.
2. The sequence $l_{id}^{(k)}$ converges if the likelihood function is bounded.

Notes:

1. Even though the sequence $l_{id}^{(k)}$ may converge, there is in general no guarantee that $\hat{\theta}^{(k)}$ converges, e.g., the EM algorithm may jump back and forth between two attractors.

2. Even if the algorithm has a stable point, there is in general no guarantee that this stable point is a global maximum of the likelihood function, or even a local maximum. For some problems, convergence to a local maximum has been demonstrated.

3. The solution in general depends on the initialization.

4. Several variants of the EM algorithm have been developed. One of the most promising ones is Fessler and Heros SAGE algorithm (space-alternating generalized EM) whose convergence rate is often far superior to that of the EM algorithm [7].

5. There is a close relationship between convergence rate of the EM algorithm and Fisher information, as described in [5, 7].

6 Maximum A Posteriori Estimation

In this section, we assume that a prior is available for the original image $\mathbf{x}$. Assume for convenience that $\mathbf{x}$ is defined over a finite domain and is a random N-vector with pdf $p(\mathbf{x})$. By Bayes' rule, the posterior pdf for $\mathbf{x}$ is given by

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}. \tag{9}$$

The ML estimator seeks $\mathbf{x}$ that maximizes the likelihood $p(\mathbf{y}|\mathbf{x})$. Similarly, the Maximum A Posteriori (MAP) estimator seeks $\mathbf{x}$ that maximizes the posterior likelihood $p(\mathbf{x}|\mathbf{y})$:

$$\hat{\mathbf{x}}_{MAP} = \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}). \tag{10}$$

This is in keeping with the intuitive notion that $p(\mathbf{x})$ reflects our prior belief about $\mathbf{x}$, and that $p(\mathbf{x}|\mathbf{y})$ reflects our updated belief after observing the data $\mathbf{y}$.

Also note from (9) and (10) that

$$\hat{\mathbf{x}}_{MAP} = \arg\max_{x} [p(\mathbf{y}|\mathbf{x})p(\mathbf{x})],$$

i.e., the normalizing factor $p(\mathbf{y})$ in (9) plays no role in MAP estimation.

6.1 Example

Let the observational model be given by

$$Y_n \sim \text{iid Poisson}(\theta) \quad 1 \le n \le N$$

and the prior on θ by

$$p(\theta) = a\, e^{-a\theta}\, 1_{\{\theta>0\}}.$$

We have (for $\theta > 0$)

$$\begin{aligned} \ln p(\mathbf{y}|\theta) &= \left(\sum_{n=1}^{N} y_n\right) \ln \theta - N\theta - \sum_{n=1}^{N} \ln y_n! \\ \ln p(\theta) &= -a\theta + \ln a. \end{aligned}$$

The likelihood equation is

$$0 = \frac{d \ln p(\mathbf{y}|\theta)}{d\theta} = \frac{1}{\theta} \sum_{n=1}^{N} y_n - N$$

which yields

$$\hat{\theta}_{ML} = \frac{1}{N} \sum_{n=1}^{N} y_n.$$

The MAP equation is

$$0 = \frac{d[\ln p(\mathbf{y}|\theta) + \ln p(\theta)]}{d\theta} = \frac{1}{\theta} \sum_{n=1}^{N} y_n - N - a$$

which yields

$$\hat{\theta}_{MAP} = \frac{1}{N + a} \sum_{n=1}^{N} y_n.$$

In this example $\hat{\theta}_{MAP}$ is smaller than $\hat{\theta}_{ML}$, but the difference vanishes as $N \to \infty$. At the opposite extreme, if $N = 1$ and $a \gg 1$, the prior is very sharply peaked in the vicinity of 0, while the likelihood function is relatively flat when $N = 1$. In this case, the MAP estimator does not trust the data. This is fine if the prior is correct, but also hints that misspecification of the prior can lead to disaster in some cases.

6.2 Cramer-Rao Bound

For unbiased estimators of $\mathbf{x}$, we have

$$\mathrm{Cov}(\hat{\mathbf{X}}) \geq \mathsf{I}_{\mathbf{X}}^{-1} \tag{11}$$

where

$$\mathsf{I}_{\mathbf{X}} = \mathsf{I}_{\mathbf{X}}^{d} + \mathsf{I}_{\mathbf{X}}^{p} \tag{12}$$

and

$$\begin{aligned} \mathsf{I}_{\mathbf{X}}^{d} &= \mathbb{E}[(\nabla_{\mathbf{X}} \ln p(\mathbf{Y}|\mathbf{X}))(\nabla_{\mathbf{X}} \ln p(\mathbf{Y}|\mathbf{X}))^T] \\ \mathsf{I}_{\mathbf{X}}^{p} &= \mathbb{E}[(\nabla_{\mathbf{X}} \ln p(\mathbf{X}))(\nabla_{\mathbf{X}} \ln p(\mathbf{X}))^T] \end{aligned}$$

are (under regularity conditions) the Fisher information matrices corresponding to the data and the prior, respectively.

7 Computation of MAP estimates

In this section, we shall present two popular computational methods for computing MAP estimates, i.e., finding a maximizer $\hat{\mathbf{x}}_{MAP}$ of the posterior distribution $p(\mathbf{x}|\mathbf{y})$. Note that

- The MAP estimate is evaluated as the solution to the equivalent maximization problem

$$\max_{\mathbf{x}} \{\ln p(\mathbf{y}|\mathbf{x}) + \ln p(\mathbf{x})\}. \tag{13}$$

- If the maximizer is not unique, any of these maximizers can be called MAP estimate.
- The cost function may have one or more local maxima which are not global maxima and are therefore not MAP estimates.

7.1 Convex Optimization

In many image restoration problems, both the log likeilhood and the log prior terms in (13) are concave. This presents several important advantages:

1. Any local maximum of (13) is a global maximum.
2. Established convex programming algorithms can be used to compute the solution.

It is however possible that off-the-shelf convex programming algorithms fail to exploit some specific structure in the cost function. Also, the large dimensionality of $\mathbf{x}$ (say hundreds of thousands of pixels) may preclude the use of some standard algorithms. For this reason, even when the optimization problem is convex, tailor-made solutions such as the one introduced below are popular.

7.2 ICM Algorithm

The Iterated Conditioned Modes (ICM) algorithm was introduced by Besag [9]. The ICM algorithm breaks down the original high-dimensional optimization problem into a sequence of simpler one-dimensional optimization problems, involving one single variable at a time. The algorithm is computationally attractive under priors of the form

$$\ln p(\mathbf{x}) = -\sum_{i \sim j} \varphi(x_i - x_j)$$

(Gibbs random field models), where $\sum_{i \sim j}$ denotes summation over i and j that are in the same clique. The ICM algorithm exploits the local nature of the prior. For the algorithm to

be efficient, a local change in the image $\mathbf{x}$ should only have a local influence in the data $\mathbf{y}$. Consider the case where the observational model involves independent noise:

$$p(\mathbf{y}|\mathbf{x}) = \exp\left\{\sum_i \ln p(y_i|x_i)\right\}. \tag{14}$$

The algorithm works as follows. Fix all pixels x_j except for $j = i$. Maximize the MAP criterion over x_i only:

$$\hat{x}_i^{(k+1)} = \arg\max_{x_i}\left[\ln p(y_i|x_i) - \sum_{i\sim j}\varphi(x_i - \hat{x}_j^{(k)})\right].$$

Then repeat for the next pixel and so on, successively optimizing the MAP criterion over individual pixels. ICM is therefore a coordinate-descent optimization algorithm. The algorithm is greedy, as it optimizes only one variable at a time without considering the effect on other coordinates. The MAP criterion is nondecreasing with iteration number. Furthermore, if the MAP criterion is differentiable with respect to x, the ICM algorithm is guaranteed to converge to a local maximum. These useful properties, coupled with the computational simplicity of the algorithm, have made ICM very popular in image restoration.

Note: If the MAP criterion is nondifferentiable, the ICM algorithm need not converge to a local maximum. Specifically, the algorithm may get stuck on a ridge of the cost function. For instance, maximize

$$f(\mathbf{x}) = \begin{cases} x_1 & : \ x_0 \geq x_1 \\ x_0 & : \ \text{else.} \end{cases} \qquad \text{over } \mathbf{x} = (x_0, x_1) \in [0,1]^2.$$

The function has a maximum (equal to 1) at location $\mathbf{x} = (1,1)$. The graph of this function (see below) presents a ridge on the segment $0 \leq x_0 = x_1 \leq 1$. For any initialization $\mathbf{x}^0$, performing the maximization of f along any coordinate yields $\mathbf{x}^1$ on the ridge; subsequent optimization steps do not move the estimate, so $\mathbf{x}^k = \mathbf{x}^1$ for all $k > 1$.

7.3 Simulated Annealing

Geman and Geman [8] developed a Simulated Annealing (SA) algorithm which is guaranteed to converge to the global maximum of the MAP criterion under mild conditions. Assume that the posterior density can be written in the form of a Gibbs distribution,

$$p(\mathbf{x}|\mathbf{y}) \propto \exp\{-U(\mathbf{x})\}. \tag{15}$$

(Here $U(\mathbf{x})$ depends on $\mathbf{y}$.) For instance, assume the prior for $\mathbf{x}$ is a Gibbs distribution of the form

$$p(\mathbf{x}) \propto \exp\left\{-\sum_i \psi(x_i) - \sum_{i\sim j}\varphi(x_i - x_j)\right\}$$

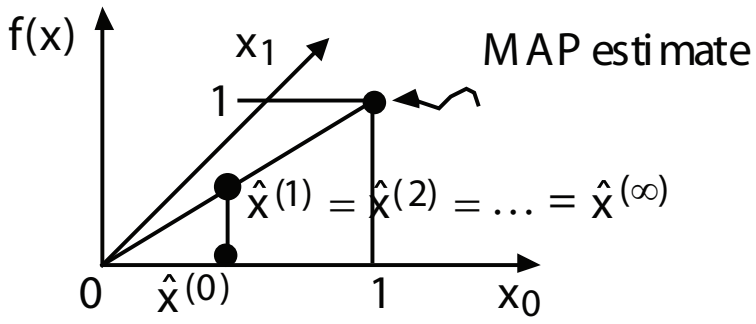


Figure 9: ICM getting stuck on a ridge.

and the observational model involves independent noise as in (14). The posterior $p(\mathbf{x}|\mathbf{y})$ is therefore of the form (15) with

$$U(\mathbf{x}) = \sum_i [-\ln p(y_i|x_i) + \psi(x_i)] + \sum_{i\sim j} \varphi(x_i - x_j).$$

For the SA approach to be computationally efficient, the neighborhoods of the MRF $p(\mathbf{x}|\mathbf{y})$ should be relatively small. Now let

$$p_T(\mathbf{x}|\mathbf{y}) \triangleq \frac{1}{Z_T} \exp\left\{-\frac{U(\mathbf{x})}{T}\right\}.$$

The maxima of both $p(\mathbf{x}|\mathbf{y})$ and $p_T(\mathbf{x}|\mathbf{y})$ are minima of $U(\mathbf{x})$. As $T \to 0$, the distribution $p_T(\mathbf{x}|\mathbf{y})$ gets more peaked in the vicinity of

$$\hat{\mathbf{x}}_{MAP} = \arg\min_{\mathbf{x}} U(\mathbf{x}).$$

The SA algorithm uses a sampling method (e.g., Gibbs or Metropolis) to generate a sample $\mathbf{x}$ from $p_T(\cdot|\mathbf{y})$. At first sight we would just need to choose an arbitrarily low temperature $T \ll 1$, use the Gibbs sampler, and obtain a solution that is almost surely close to $\hat{\mathbf{x}}_{MAP}$. Unfortunately the Gibbs sampler converges extremely slowly at low temperatures, so this method is generally considered unpractical.

A better idea is to initially choose a moderate value for T and apply the Gibbs sampler, letting T slowly decrease to 0 during the iterations. Denoting the temperature at iteration k of the Gibbs sampler by T_k, the question is now to choose an appropriate *cooling schedule* T_k. Geman and Geman showed that if

$$T_k = \frac{T}{1 + \log k}, \tag{16}$$

the SA algorithm converges to $\hat{\mathbf{x}}_{MAP}$ almost surely.

The advantage of the method is that it is guaranteed to converge to the global maximum of the MAP criterion. The main disadvantage is that the method is computationally very

intensive because the logarithmic cooling schedule (16) is very slow. One could use a faster cooling schedule and still possibly obtain good results, but there is no guarantee this will happen.

For that reason, in practice, simpler and fast algorithms such as ICM are often preferred to SA. However, SA can be used to assess the performance loss due to the use of suboptimal numerical algorithms.

8 Regularization Methods

Let us briefly digress from our study of statistical image restoration methods to introduce an important class of deterministic image restoration methods. Consider the inverse-problem setup of Model 4, where H is possibly a nonlinear operator.

As discussed earlier, this setup includes problems of deblurring (where H is LSI), inversion of the Radon transform (where H is linear but not shift-invariant), etc. In such problems, the inverse filter solution is nonunique or unstable under even very small amounts of noise corrupting measurements $\mathbf{y}$. In particular, in deblurring problems, the inverse filter solution is corrupted by large amounts of high-frequency noise.

The method of regularization stabilizes such solutions. The regularized estimator is defined as follows:

$$\hat{\mathbf{x}}_{REG} = \arg\min_{\mathbf{x}} \{\mathcal{E}(\mathbf{x}) = \|\mathbf{y} - H\mathbf{x}\|^2 + \lambda\phi(\mathbf{x})\} \tag{17}$$

where $\lambda \geq 0$ is the regularization parameter, and $\phi(\mathbf{x})$ is the regularization functional. The dimensionalities of $\mathbf{x}$ and $\mathbf{y}$ may be different.

The regularization functional $\phi(\mathbf{x})$ is chosen so as to penalize undesirable solutions. The regularization parameter λ determines the amount of regularization. For $\lambda \to 0$, the solution tends to the inverse filter solution. For $\lambda \to \infty$, the estimator essentially tries to minimize $\phi(\mathbf{x})$.

Example 1: Continuous coordinates $t = (t_1, t_2)$: a popular choice is

$$\phi(\mathbf{x}) = \int \|\nabla x(t)\|^2 \, dt = \int \left|\frac{\partial x(t)}{\partial t_1}\right|^2 + \left|\frac{\partial x(t)}{\partial t_2}\right|^2 \, dt$$

which penalizes large gradients (Tikhonov regularization).

Example 2: Quadratic penalty. Here

$$\phi(\mathbf{x}) = \|C\mathbf{x}\|^2 \tag{18}$$

where C represents a highpass filtering operation such as discrete differentiation. A simple 1-D discrete differentiator is

$$(C\mathbf{x})(n) = x(n) - x(n-1), \quad 1 \leq n \leq N,$$

with the convention $x(0) = 0$. In 2-D, we could define C as the concatenation of C_H and C_V, such that $\|C\mathbf{x}\|^2 = \|C_H\mathbf{x}\|^2 + \|C_V\mathbf{x}\|^2$ and

$$\begin{aligned}(C_H\mathbf{x})(n) &= x(n_1, n_2) - x(n_1 - 1, n_2)\\ (C_V\mathbf{x})(n) &= x(n_1, n_2) - x(n_1, n_2 - 1), \quad 1 \le n_1 \le N_1,\ 1 \le n_2 \le N_2,\end{aligned}$$

are respectively the arrays of horizontal and vertical discrete derivatives.

When C is a highpass filter, $\phi(\mathbf{x})$ is often referred to as a roughness penalty.

When the data fidelity term and the regularization functional are quadratic, $\hat{\mathbf{x}}_{REG}$ is the solution to a linear system. From (17) and (18), we obtain

$$0 = \nabla\mathcal{E}(\mathbf{x}) = H^T(H\mathbf{x} - \mathbf{y}) + \lambda C^T C\mathbf{x}$$

hence

$$\hat{\mathbf{x}}_{REG} = (H^T H + \lambda C^T C)^{-1} H^T \mathbf{y}$$

Relation to MAP Estimation. Note that the regularized estimator of (17) may be written in the equivalent form

$$\hat{\mathbf{x}}_{REG} = \arg\max_{\mathbf{x}} \left\{ -\frac{1}{2\sigma^2}\|\mathbf{y} - H\mathbf{x}\|^2 - \frac{\lambda}{2\sigma^2}\phi(\mathbf{x}) \right\}. \tag{19}$$

Hence regularization is equivalent to MAP estimation using the Gaussian measurement model $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(H\mathbf{x}, \sigma^2 I)$ and the prior

$$p(\mathbf{x}) = \frac{1}{Z} \exp\left\{ -\frac{\lambda}{2\sigma^2}\phi(\mathbf{x}) \right\},$$

provided that the exponential term is integrable:

$$Z = \int \exp\left\{ -\frac{\lambda}{2\sigma^2}\phi(\mathbf{x}) \right\} d\mathbf{x} < \infty. \tag{20}$$

The regularization parameter λ controls the peakness of prior.

In summary, there exists a direct equivalence between the deterministic and statistical approaches, provided the integrability condition (20) holds. Even if (20) does not hold, one may still view (19) as a MAP estimation criterion in which $p(\mathbf{x})$ is an *improper prior* (which is not integrable) [11]. The equivalence between the deterministic and statistical approaches van be useful for designing a suitable $\phi(\mathbf{x})$.

9 Constrained Optimization

This is another class of deterministic image restoration methods which was already used in the sixties. Consider the same setup as in Sec. 8. Here we assume that some information about the errors corrupting the measurements is available. Specifically, assume the mean-squared value

$$\|\mathbf{y} - H\mathbf{x}\|^2 = N\sigma^2 \tag{21}$$

of the errors is known. Define a cost function $\phi(\mathbf{x})$ which penalizes undesirable (e.g., rough) images. The constrained estimate is defined as

$$\hat{\mathbf{x}}_{CSTR} = \arg\min_{\mathbf{x}} \phi(\mathbf{x}) \tag{22}$$

where the minimization is subject to the constraint (21).

We can write the constrained optimization problem as a Lagrange optimization problem:

$$\hat{\mathbf{x}}_{CSTR} = \arg\min_{\mathbf{x}} \left[\phi(\mathbf{x}) + \mu\|\mathbf{y} - H\mathbf{x}\|^2\right] \tag{23}$$

where the Lagrange multiplier $\mu \geq 0$ is chosen so as to satisfy the constraint equation (21).

From the Lagrangian formulation of the constrained optimization problem we see the equivalence with the regularization method of Section 8. Specifically, $\phi(\mathbf{x})$ plays the role of a regularization functional, and the value of the regularization parameter is $\lambda = \mu^{-1}$. The constraint (21) determines the value of the regularization parameter.

Note 1: If $\phi(\mathbf{x})$ is quadratic in $\mathbf{x}$, the estimation method (22) is known as constrained least-squares.

Note 2: Again this is a purely deterministic method. However, if it is known that the measurement errors $\mathbf{y} - H\mathbf{x}$ are random and iid with mean zero and variance σ^2, then by the Central Limit Theorem, we have $\|\mathbf{y} - H\mathbf{x}\|^2 \to N\sigma^2$ almost surely as $N \to \infty$. This is an often cited "justification" for the method of constrained optimization under such limited knowledge of the statistics of the measurement errors. However, imposing the condition that the estimator $\hat{\mathbf{x}}$ satisfy the condition $\|\mathbf{y} - H\hat{\mathbf{x}}\|^2 = N\sigma^2$ may be a rather questionable strategy. For instance, if $\mathbf{X}$ is iid $\mathcal{N}(0, \sigma_X^2)$, H is the $N \times N$ identity matrix, and $\mathbf{W} = \mathbf{Y} - H\mathbf{X}$ is iid $\mathcal{N}(0, \sigma^2)$, the Wiener filter solution

$$\hat{\mathbf{x}} = \frac{\sigma_X^2}{\sigma_X^2 + \sigma^2}\mathbf{x}$$

is optimal in the MSE sense, and for this choice we have

$$\mathbb{E}\|\mathbf{Y} - \hat{\mathbf{X}}\|^2 = N\sigma^2 \frac{\sigma_X^2}{\sigma_X^2 + \sigma^2} < N\sigma^2.$$

The trivial estimator $\hat{\mathbf{x}} = \mathbf{y}$, on the other hand, satisfies (21) but can hardly be viewed as a good estimator.

10 Choice of Regularization Functional

In this section, we discuss criteria for choosing the regularization functional $\phi(\mathbf{x})$. As discussed in Sec. 8, $\phi(\mathbf{x})$ should capture a priori knowledge about the image. In particular, $\phi(\mathbf{x})$ should capture the smoothness of homogeneous areas of the image while allowing sharp transitions between different regions. Finally, computational considerations suggest that $\phi(\mathbf{x})$ should be convex, or at least should be chosen in a way that facilitates numerical optimization of the cost function $\mathcal{E}(\mathbf{x})$.

10.1 Total variation image restoration

The total variation (TV) norm of image $\mathbf{x}$ is defined as

$$\phi(\mathbf{x}) = \int \|\nabla x(t)\| \, dt.$$

This penalty is convex, but not strictly convex. It has become popular in image restoration, following work by Rudin *et al.* [13]. To see why, consider the 1-D equivalent of $\phi(\mathbf{x})$ for $\mathbf{x} = \{x(t),\, 0 \le t \le 1\}$:

$$\phi(\mathbf{x}) = \int_0^1 |x'(t)| \, dt = \int_0^1 |dx(t)|.$$

This definition can be formally generalized to the case of nondifferentiable functions by allowing Dirac impulses in $x'(t)$, or by defining

$$\phi(\mathbf{x}) = \sup \sum_i |x(t_{i+1}) - x(t_i)|$$

where the supremum is over all increasing sequences $\{t_i\}$ in the interval $[0, 1]$.

The TV norm has an interesting property. All nondecreasing functions over an interval with fixed endpoint values have the same TV norm. For instance, fixing $x(0) = 0$ and $x(1) = A$, all such functions have TV norm equal to $|x(1) - x(0)| = A$; see Fig. 10 for an illustration.

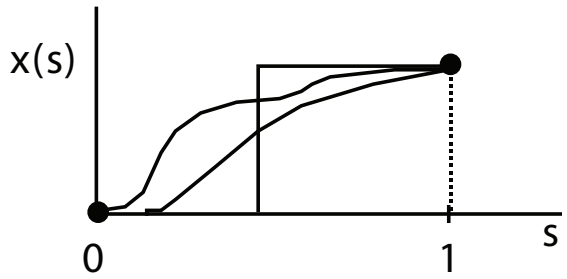


Figure 10: All three functions above have the same TV norm.

In particular, note that the TV norm does not penalize edges. TV regularization has been successfully employed to denoise image with sharp edges [13].

10.2 Concavity and edge-preservation property

We expand the analysis above and consider regularization penalties of the form [14, 16]

$$\phi(\mathbf{x}) = \sum_{i=1}^{N} \varphi(x_i - x_{i-1})$$

where $\varphi(\cdot)$ is an even function, and $\mathbf{x}$ is a discrete 1-D signal. To see what types of configuration $\phi(\mathbf{x})$ favors, assume that $x_0 = 0$ and $x_N = A$ are fixed. Then choose $x_1, \cdots, x_{N-1}$ that minimize $\phi(\mathbf{x})$. The solution strongly depends on the convexity or concavity properties of $\varphi(\cdot)$.

Case I: $\varphi(\cdot)$ is strictly convex. Then

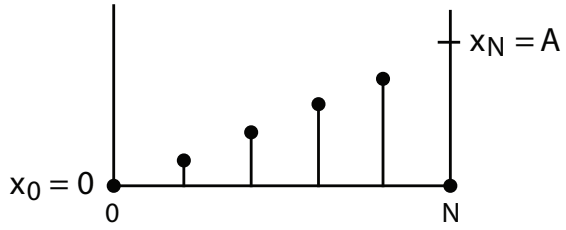
$$\frac{1}{N}\phi(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\varphi(x_i - x_{i-1}) \geq \varphi\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - x_{i-1})\right) = \varphi\left(\frac{A}{N}\right)$$

with equality if and only if

$$x_i = \frac{Ai}{N}, \quad 1 \leq i \leq N-1. \tag{24}$$

The solution is the linear interpolator between x_0 and x_N.

Note: if $\varphi(\cdot)$ is simply convex but not strictly convex, the condition (24) is sufficient but might not be necessary for optimality.



Case II: $\varphi(\cdot)$ is strictly concave over $\mathbb{R}^+$. Then any bipolar sequence with values 0 and A and a single jump at some $i^* \in \{0, 1, \cdots, N\}$ is optimal, as illustrated below. The proof is left as an exercise.

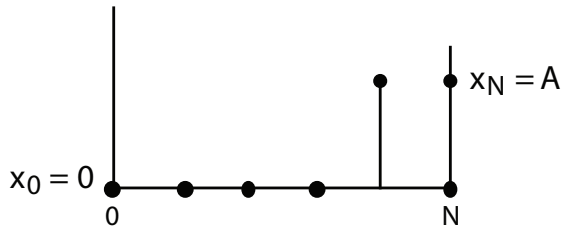


Figure 11: Function with a jump at $i^* = N - 1$.

Commonly used strictly concave functions include $\varphi(u) = |u|^p$ with $0 < p < 1$ (Generalized Gaussian MRFs).

Case III: The choice $\varphi(u) = |u|$ corresponds to the TV regularization method. As shown above, any nondecreasing sequence $\mathbf{x}$ is optimal in that case.

Similar results hold in two dimensions, for potential functions of the form

$$\phi(\mathbf{x}) = \sum_{i \sim j} \varphi(x_i - x_j)$$

where the summation is over all nearest neighbors (i, j).

Remarks.

1. It is remarkable that the optimal solution depends on φ only via its convexity or concavity property. For strictly convex $\varphi(\cdot)$, the estimator minimizes the disparity between the differences $x_i - x_{i-1}$ (it makes them all equal). For strictly concave $\varphi(\cdot)$, the estimator maximizes the disparity between the differences $x_i - x_{i-1}$ (they are equal to 0 or A).

2. For strictly convex $\phi(\mathbf{x})$ (hence strictly convex $\mathcal{E}(\mathbf{x})$), it can be shown that the estimator $\hat{\mathbf{x}}(\mathbf{y})$ is a continuous function of the data $\mathbf{y}$.

3. For strictly concave $\phi(\mathbf{x})$, the function $\mathcal{E}(\mathbf{x})$ has in general many local minima, and $\hat{\mathbf{x}}(\mathbf{y})$ is not continuous. The estimate is then said to be unstable (which is not necessarily undesirable, as $\hat{\mathbf{x}}(\mathbf{y})$ acts as an edge detector).

4. If $\phi(\mathbf{x})$ is strictly concave and the true signal $\mathbf{x}^*$ is an edge, then for sufficiently large regularization parameter λ, $\mathbf{x}^*$ is a local minimum of $\mathcal{E}(\mathbf{x})$ [14]. The key in the proof is that $\phi'(0^+) > 0$. This property does not hold for strictly convex $\phi(\mathbf{x})$.

The theory above shows that strictly concave (over $\mathbb{R}^+$) functions $\varphi(\cdot)$ favor the formation of constant-intensity regions separated by sharp discontinuities. This can be visually unpleasant [17]. In practice, convex potential functions that are quadratic near $u = 0$ and nearly linear as $|u| \to \infty$ have given good results. This includes the so-called Huber potential function

$$\varphi(u) = \begin{cases} \frac{u^2}{2} & : \ |u| < 1 \\ |u| - \frac{1}{2} & : \ |u| > 1 \end{cases}$$

and Green's potential function $\varphi(u) = \ln \cosh u$, which is infinitely differentiable.

10.3 Wavelet-domain priors

As discussed in Chapter III of these notes, some very good (and tractable) statistical image models have been developed in the wavelet domain. In this section we restrict our attention to wavelet models involving *independent coefficients*. Denoting by W the Discrete Wavelet Transform and by s and k the subband and the location of a wavelet coefficient within the subband respectively, we write the set of wavelet coefficients $\{\tilde{x}_{sk}\}$ as $\tilde{\mathbf{x}} = \mathsf{W}\mathbf{x}$. By our independence assumption on the wavelet coefficients, we have

$$p(\tilde{\mathbf{x}}) = \prod_{s,k} p_{sk}(\tilde{x}_{sk}).$$

Hence the negative log prior is additive over the wavelet coefficients. For the GGD wavelet models discussed in Chapter III, we have

$$\varphi_{sk}(\tilde{x}_{sk}) \triangleq -\ln p_{sk}(\tilde{x}_{sk}) = C_{sk}|\tilde{x}_{sk}|^p$$

where C_{sk} is a positive normalizing constant. When $0 < p < 1$, the function $\varphi_{sk}(u)$ is strictly concave for $u > 0$.

For denoising problems (with iid Gaussian noise), denote by $\tilde{\mathbf{y}} = \{\tilde{y}_{sk}\}$ the wavelet coefficients of the noisy image data $\mathbf{y}$. The cost function to be minimized is

$$\begin{aligned} \mathcal{E}(\mathbf{x}) &= \frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{x}\|^2 + \sum_{s,k} C_{sk}|\tilde{x}_{sk}|^p \\ &= \frac{1}{2\sigma^2}\|\tilde{\mathbf{y}}-\tilde{\mathbf{x}}\|^2 + \sum_{s,k} C_{sk}|\tilde{x}_{sk}|^p \\ &= \sum_{s,k} \mathcal{E}_{sk}(\tilde{x}_{sk}) \end{aligned} \tag{25}$$

where the second equality follows from the orthonormality of the wavelet transform, and in the third line we have defined

$$\mathcal{E}_{sk}(\tilde{x}_{sk}) = \frac{1}{2\sigma^2}(\tilde{y}_{sk}-\tilde{x}_{sk})^2 + C_{sk}|\tilde{x}_{sk}|^p \tag{26}$$

Equation (25) shows that $\mathcal{E}(\mathbf{x})$ is additive over the wavelet coefficients $\tilde{x}_{sk}$. Hence each term in the sum can be minimized independently of the other terms. The N-dimensional optimization problem can be reduced to N independent 1-D optimization problems of the form

$$\min_{\tilde{x}_{sk}} \mathcal{E}_{sk}(\tilde{x}_{sk}). \tag{27}$$

The solution to (27) is given below.

10.4 Shrinkage Estimators

In this section, we study the estimator

$$\hat{x}(y) = \arg\min_{\mathbf{x}} \left\{ \mathcal{E}(x) = \frac{1}{2\sigma^2}(x-y)^2 + \lambda\,\varphi(x) \right\}, \quad x, y \in \mathbb{R}. \tag{28}$$

We shall make the following assumptions:

(A1) $\varphi(0) = 0$.

(A2) $\varphi(x)$ is a symmetric function.

(A3) $\varphi(x)$ is nondecreasing for $x \geq 0$.

Assumption (A1) does not cause any loss of generality because $\hat{x}(y)$ is not affected by addition of constant terms in the cost function $\mathcal{E}(x)$. Assumption (A2) implies that $\hat{x}(y)$ is an antisymmetric function. Assumptions (A2) and (A3) together imply first that $\hat{x}(y)$ is nondecreasing, and second that $\hat{x}(y)$ is a *shrinkage estimator*, i.e., satisfies the property

$$|\hat{x}(y)| \leq |y|, \quad \forall y.$$

The proof of these claims is left as an exercise for the reader.

Before deriving the solution to the general problem (28), we present a few special cases that are insightful and yield estimators that have been successfully used in image denoising practice. The derivative of $\varphi(x)$, when it exists, is denoted by $\varphi'(x)$. Then the derivative of the cost function $\mathcal{E}(x)$ also exists and is given by

$$\mathcal{E}'(x) = \frac{1}{\sigma^2}(x - y) + \lambda\, \varphi'(x). \tag{29}$$

Case I: $\varphi(x) = x^2$. Then (28) is quadratic in x, and

$$\mathcal{E}'(x) = \frac{1}{\sigma^2}(x - y) + 2\lambda x$$

always exists. Setting $\mathcal{E}'(x) = 0$, we obtain the Wiener filter solution

$$\hat{x} = \frac{1}{1 + 2\lambda\sigma^2}\, y,$$

which represents a fixed attenuation of the observed noisy coefficient y.

Case II: $\varphi(x) = |x|$. Then $\varphi'(x) = sgn(x)$ for all $x \neq 0$ but $\varphi'(0)$ does not exist (the left and right derivatives of $\varphi(x)$ at $x = 0$ are not equal). There are two regimes to be considered. In the first one, the minimum of $\mathcal{E}(x)$ is achieved at some $\hat{x} \neq 0$, so $\mathcal{E}'(\hat{x})$ exists. In the second regime, $\hat{x} = 0$. Verify that the solution takes the form

$$\hat{x} = \begin{cases} y - 2\sigma^2\lambda & : \ y > 2\sigma^2\lambda \\ 0 & : \ |y| \leq 2\sigma^2\lambda \\ y + 2\sigma^2\lambda & : \ y < -2\sigma^2\lambda \end{cases}$$

or for short,

$$\hat{x} = (|y| - 2\sigma^2\lambda)^+ sgn(y) \tag{30}$$

where $(u)^+ \triangleq \max(u, 0)$. The estimate is obtained by applying a *soft threshold* to y: $\hat{x}$ is set to zero for all $|y|$ smaller than the threshold $2\sigma^2\lambda$.

Case III: $\varphi(x) = 1_{\{x \neq 0\}}$. This is sometimes called a *complexity prior*: a fixed cost is attached to any nonzero coefficient. Again there are two regimes to be considered:

- If $\hat{x} = 0$, then $\mathcal{E}(\hat{x}) = \frac{1}{2\sigma^2} y^2$.

- If $\hat{x} \neq 0$, then the minimizing $\hat{x}$ must be equal to y, and $\mathcal{E}(\hat{x}) = \lambda$.

The only consistent solution is therefore a *hard threshold* on y:

$$\hat{x} = y\, 1_{\{|y| > \sqrt{2\sigma^2\lambda}\}}. \tag{31}$$

Case IV: $\varphi(x)$ is strictly concave for $x \geq 0$, and the right derivative of $\varphi(x)$ at $x = 0$ is strictly positive:

$$\varphi'(0^+) > 0.$$

In this case one can demonstrate the existence of a threshold effect [20] with threshold equal to $\sigma^2\varphi'(0^+)$. The estimator $\hat{x}(y)$ possesses the following properties:

- $\hat{x}(y) = 0$ iff $|y| \leq \sigma^2\lambda\varphi'(0^+)$.
- The right derivative of the function $\hat{x}(y)$ at $y = \sigma^2\lambda\varphi'(0^+)$ is infinite.
- $\hat{x}(y) \to y$ as $y \to \infty$.

The above conditions on $\varphi(x)$ apply to the family of penalties $\varphi(x) = |x|^p$ when $0 < p < 1$. As $p \to 1$, the estimator tends (pointwise, i.e., for any fixed y) to the soft-threshold estimator of (30). As $p \to 0$, the estimator tends (pointwise) to the hard-threshold estimator of (31).

Returning to the original N-dimensional optimization problem of (25), we conclude that despite the nonconvexity of $\mathcal{E}(\mathbf{x})$, the solution is easy to compute because it can be decomposed into N independent optimization problems, each of which is nonconvex but is easy to solve. When $\varphi(\cdot)$ is strictly concave over $\mathbb{R}^+$, we note the appearance of thresholding effects, which favor sparse estimates (in which many estimated coefficients are zero).

In the presence of blur, it seems much harder to obtain the globally optimal solution. One approach is to compute $\hat{\mathbf{x}}(\mathbf{y})$ numerically, using an ICM algorithm [19]. This approach has become quite popular in recent image restoration literature. Still many authors prefer to choose a convex $\phi(\mathbf{x})$, so the resulting minimization problem remains convex. The paper [18] presents an approximate solution to the deblurring problem using $\varphi(u) \propto |u|$ and a conjugate gradient method. Results show noticeable visual improvements over the TV restoration method.

11 Set-Theoretic Estimation

In some applications, prior information about the signal $\mathbf{x}$ is not statistical but takes the form of *membership to a set* [21]. For instance, $\mathbf{x}$ may belong to

$$\mathcal{C} \triangleq \cap_{i=1}^{m} \mathcal{C}_i \tag{32}$$

where $m = 5$ and

$$\begin{aligned}
\mathcal{C}_1 &= \left\{\mathbf{x} \,:\, \|\mathbf{y} - H\mathbf{x}\|^2 \le N\sigma^2\right\}, \\
\mathcal{C}_2 &= \left\{\mathbf{x} \,:\, \|C\mathbf{x}\|^2 \le N\mu_2\right\}, \\
\mathcal{C}_3 &= \left\{\mathbf{x} \,:\, \sum_{i=1}^{N} |x_i| \le N\mu_3\right\}, \\
\mathcal{C}_4 &= \left\{\mathbf{x} \,:\, x_1 = x_2 = 0\right\}, \\
\mathcal{C}_5 &= \left\{\mathbf{x} \,:\, (\mathscr{F}\mathbf{x})_i = 0,\ N/2 \le i \le N\right\}.
\end{aligned} \tag{33}$$

In the last line, $\mathscr{F}$ denotes the N-point DFT. The goal of set-theoretic estimation is simply to produce some $\mathbf{x} \in \mathcal{C}$. There is no notion of optimality other than membership in $\mathcal{C}$. This principle is rather different from the principle of *point estimation* espounded so far, where the goal is to produce a solution that optimizes some criterion (e.g., ML, MAP, or Bayes).

For mathematical convenience we shall assume that the sets $\mathcal{C}_i$ are convex, i.e., if $\mathbf{x}^1$ and $\mathbf{x}^2$ are in $\mathcal{C}_i$, then so is $\lambda\mathbf{x}^1 + (1-\lambda)\mathbf{x}^2$ for all $\lambda \in [0,1]$. It will also be assumed that the intersection of the sets $\mathcal{C}_i$ is nonempty.

If the sets $\mathcal{C}_i$ are convex and $\mathcal{C} \neq \emptyset$, a Projection Onto Convex Sets (POCS) algorithm may be used to produce a sequence of estimates that converge to an element of $\mathcal{C}$. The estimate $\hat{\mathbf{x}}^{(k)}$ at iteration k of the algorithm is obtained by projecting the previous estimate $\hat{\mathbf{x}}^{(k-1)}$ onto one of the constraint sets $\mathcal{C}_i$. Note that in general the limit $\lim_{k\to\infty} \hat{\mathbf{x}}^{(k)}$ need not be the projection of the initial estimate $\hat{\mathbf{x}}^{(0)}$ onto $\mathcal{C}$.

Fig. 12 illustrates POCS in the case of two sets $\mathcal{C}_1$ and $\mathcal{C}_2$. Let the initial estimate be $\hat{\mathbf{x}}^{(0)} = \mathbf{y}_2$, and the algorithm perform successive projections onto $\mathcal{C}_2$, $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_1$, etc. The algorithm converges to $\hat{\mathbf{x}}_{2,POCS} \in \mathcal{C}_1 \cap \mathcal{C}_2$, which in this case coincides with the projection of $\mathbf{y}_2$ onto $\mathcal{C}$.

Note that a different sequence of estimates (and possibly a different limit) is obtained if the order of the projections is changed. This is clearly seen from Fig. 12 when the initial estimate is $\hat{\mathbf{x}}^{(0)} = \mathbf{y}_1$. Projecting first onto $\mathcal{C}_1$ and next onto $\mathcal{C}_2$ yields the final solution $\hat{\mathbf{x}}_{1,POCS} \in \mathcal{C}_1 \cap \mathcal{C}_2$ after only two steps of the algorithm. A different solution $\hat{\mathbf{x}}_{CML}$ (the acronym will be explained in Sec. 11.2) is obtained if we project first onto $\mathcal{C}_2$ and next onto $\mathcal{C}_1$. For the time being, note that $\hat{\mathbf{x}}_{CML}$ is the projection of $\mathbf{y}_1$ onto $\mathcal{C}$. From the set-theoretic viewpoint, the solutions $\hat{\mathbf{x}}_{POCS}$ and $\hat{\mathbf{x}}_{CML}$ are equally good because both belong to the desired set $\mathcal{C}$.

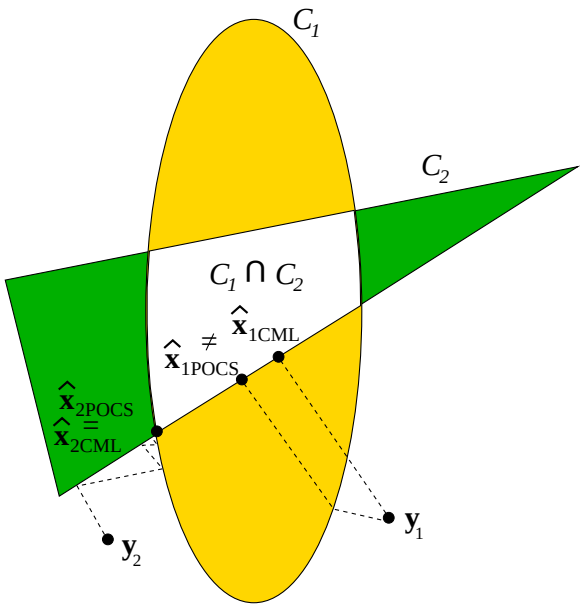


Figure 12: Successive projections using POCS method.

11.1 Projections

Let us consider the projection operations in some detail. The convex sets $\mathcal{C}_i$ can be represented in the form

$$\mathcal{C}_i = \{\mathbf{x} \ : \ \phi_i(\mathbf{x}) \leq \mu_i\}, \quad 1 \leq i \leq m \tag{34}$$

where the functions $\phi_i(\cdot)$ are convex. [1]

The projection of $\hat{\mathbf{x}}^{(k-1)}$ onto $\mathcal{C}_i$ is the solution to the minimum-distance problem

$$\min_{\mathbf{x} \in \mathcal{C}_i} \|\hat{\mathbf{x}}^{(k-1)} - \mathbf{x}\|^2$$

or equivalently,

$$\min_{\mathbf{x}} \|\hat{\mathbf{x}}^{(k-1)} - \mathbf{x}\|^2 \quad \text{subject to} \quad \phi_i(\mathbf{x}) \leq \mu_i.$$

If $\hat{\mathbf{x}}^{(k-1)}$ is already in $\mathcal{C}_i$, the set constraint is inactive, and we set $\hat{\mathbf{x}}^{(k)} = \hat{\mathbf{x}}^{(k-1)}$. If $\hat{\mathbf{x}}^{(k-1)} \notin \mathcal{C}_i$, the constraint is active and the problem can be cast as a Lagrange optimization problem,

$$\hat{\mathbf{x}}^{(k)} = \arg\min_{\mathbf{x}} \left\{ \|\hat{\mathbf{x}}^{(k-1)} - \mathbf{x}\|^2 + \lambda \phi_i(\mathbf{x}) \right\} \tag{35}$$

[1] The choice of $\phi_i(\cdot)$ and μ_i is nonunique. For instance, any convex set constraint may be written in the form $\phi_i(\mathbf{x}) \leq 0$ by choosing $\phi_i(\mathbf{x}) = 0$ for $\mathbf{x} \in \mathcal{C}_i$ and $\phi_i(\mathbf{x}) = \infty$ for $\mathbf{x} \notin \mathcal{C}_i$. Another choice is the representation of $\mathcal{C}_i$ via the Minkowski functional. For the example sets $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ of (33), there exists an obvious, natural choice for $\phi_i(\cdot)$ and μ_i. For the set $\mathcal{C}_4$, we may choose $\phi_4(\mathbf{x}) = x_1^2 + x_2^2$ or $\phi_4(\mathbf{x}) = \max(|x_1|, |x_2|)$, and $\mu_4 = 0$. Likewise, for $\mathcal{C}_5$, we may choose $\phi_5(\mathbf{x}) = \sum_{i=N/2}^{N} |(\mathscr{F}\mathbf{x})_i|^2$ or $\phi_5(\mathbf{x}) = \max_{N/2 \leq i \leq N} |(\mathscr{F}\mathbf{x})_i|$, and $\mu_5 = 0$.

where the Lagrange multiplier λ is chosen such that $\phi_i(\mathbf{x}^{(k)}) = \mu_i$. These problems are mathematically similar to those encountered in Sec. 8; the squared distance $\|\hat{\mathbf{x}}^{(k-1)} - \mathbf{x}\|^2$ of (35) replaces the data-fidelity term $\|\mathbf{y} - H\mathbf{x}\|^2$ of (17). Depending on the choice of $\phi_i(\cdot)$, the solution to (35) may be straightforward, e.g., in the case of the example sets $\mathcal{C}_1, \cdots, \mathcal{C}_5$ of (33), we respectively obtain

$$\begin{aligned}
\hat{\mathbf{x}}^{(k)} &= (I_N + \lambda H^T H)^{-1} (\hat{\mathbf{x}}^{(k-1)} + \lambda H^T \mathbf{y}), \\
\hat{\mathbf{x}}^{(k)} &= (I_N + \lambda C^T C)^{-1} \hat{\mathbf{x}}^{(k-1)}, \\
\hat{x}_n^{(k)} &= (\hat{x}_n^{(k-1)} - \lambda)^+ \, sgn(\hat{x}_n^{(k-1)}), \quad 1 \le n \le N, \\
\hat{x}_n^{(k)} &= \begin{cases} 0 & : \ n = 1, 2 \\ \hat{x}_n^{(k-1)} & : \ 3 \le n \le N, \end{cases} \\
(\mathscr{F}\hat{\mathbf{x}}^{(k)})_n &= \begin{cases} (\mathscr{F}\hat{\mathbf{x}}^{(k-1)})_n & : \ 1 \le n < N/2 \\ 0 & : \ N/2 \le n \le N \end{cases}
\end{aligned}$$

(Prove this as an exercise.)

11.2 Constrained Maximum-Likelihood

The functional representation (34) of the sets $\mathcal{C}_i$, $1 \le i \le m$, suggests a connection between set-theoretic estimation, regularization methods, and the constrained optimization methods of Sec. 9.

For instance, assume Model 2, linear blur plus additive white Gaussian noise ($\mathbf{y} = H\mathbf{x} + \mathbf{w}$), and seek the ML estimator of $\mathbf{x}$ subject to the constraints

$$\phi_i(\mathbf{x}) \le \mu_i, \quad 2 \le i \le m.$$

Equivalently, the constrained ML estimator produces $\mathbf{x}$ that minimizes $\|\mathbf{y} - H\mathbf{x}\|^2$ subject to the above $m - 1$ constraints. Fig. 12 illustrates the solution $\hat{\mathbf{x}}_{1,CML}$ obtained when $H = I_N$ and $\mathbf{y} = \mathbf{y}_1$: the solution is the projection of $\mathbf{y}$ onto $\mathcal{C}$.

The set-theoretic problem seeks some $\mathbf{x}$ that satisfies the m constraints

$$\begin{aligned}
\|\mathbf{y} - H\mathbf{x}\|^2 &\le N\sigma^2 \\
\phi_i(\mathbf{x}) &\le \mu_i, \quad 2 \le i \le m.
\end{aligned}$$

As expected and illustrated in Fig. 12, the Constrained ML and POCS solutions do not necessarily coincide.

12 Applications

12.1 Restoration of Compressed Images

Most images and video are transmitted or stored lossy compressed formats. Some of these compression artifacts (e.g., blocking artifacts in JPEG) are quite visible. Reducing them may be thought of as an image restoration problem. The image degradation model here is

$$\mathbf{y} = Q(\mathbf{x})$$

where Q is the quantization operation mapping the original image $\mathbf{x}$ into a compressed image $\mathbf{y}$ with a limited number of possible values. Hence Q is a deterministic many-to-one mapping. For instance, in the case of scalar quantization, each quantization bin is mapped into one single reproduction level. For a N-dimensional quantizer, each cell is a convex region in $\mathbb{R}^N$.

The posterior density for $\mathbf{x}$ given $\mathbf{y}$ is given by

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x})\delta(\mathbf{y} - Q(\mathbf{x})).$$

Hence

$$p(\mathbf{x}|\mathbf{y}) = \frac{1}{Z} p(\mathbf{x})\, 1_{\{\mathbf{x} \in Q^{-1}\mathbf{y}\}}$$

where $Q^{-1}(\mathbf{y})$ denotes the set of all $\mathbf{x}$ such that $Q\mathbf{x} = \mathbf{y}$, and $Z = \int_{Q^{-1}(\mathbf{y})} p(\mathbf{x})\, d\mathbf{x}$ is the normalization constant. In particular, the MAP criterion takes the form

$$\hat{\mathbf{x}}_{MAP} = \arg \max_{\mathbf{x} \in Q^{-1}(\mathbf{y})} p(\mathbf{x}).$$

Examples of such research may be found in [25, 26]. These techniques are also known as "postprocessing of compressed images."

Other types of impairment are due to data transmission. For instance, blocks of image data may be lost due to packet loss in ATM networks over which the image is transmitted [27]. Assuming the receiver knows which packets have been lost, the degradation model mapping the original $\mathbf{x}$ to the received $\mathbf{y}$ is again deterministic and many-to-one.

12.2 Restoration of Halftone Images

One may think of the binary images resulting from halftoning as a compressed version of the original continuous-tone image data. The restoration of a continuous-tone image from a halftone is called *inverse halftoning*. Special algorithms tailored to the specific nature of the halftoning process have been developed. For instance, Stevenson [28] assumes a Gauss MRF for $\mathbf{x}$ and maximizes $p(\mathbf{x})$ over $\mathbf{x} \in Q^{-1}(\mathbf{y})$.

12.3 Blind Image Restoration

Consider the deblurring model #2 in which the blurring filter H is a function of some unknown parameter θ, see Fig. 13. For the Gaussian filters that are often used to model optical blur, θ could represent the standard deviation of the Gaussian. For motion blurs, θ would represent the motion vector.

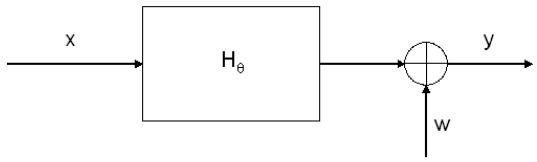


Figure 13: Partially unknown blurring filter.

We would now like to estimate both the image $\mathbf{x}$ and the parameter θ. In some cases, θ may be viewed as a nuisance parameter which does not need to be estimated accurately. However, in modern applications of image restoration, including forensics, accurate estimation of θ is useful.

The functional dependence of H on θ is critical. In some cases, the image $\mathbf{x}$ cannot be reconstructed without ambiguity. For instance, if θ is a scale factor for the filter H, then $\mathbf{x}$ can be at best reconstructed up to an arbitrary multiplying factor.

A general model for problems of the above kind is shown in Fig. 14. The measurement model $p(\mathbf{y}|\mathbf{x}, \theta)$ depends on unknown parameter θ.

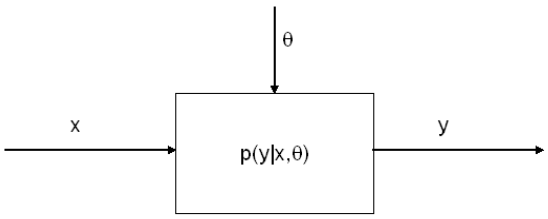


Figure 14: General model for blind image restoration.

Several approaches may be used:

1. If θ is treated as an unknown deterministic parameter, then one can first estimate θ using the method of maximum likelihood:

$$\hat{\theta}_{ML} = \arg\max_{\theta} \int p(\mathbf{y}|\mathbf{x}, \theta) p(\mathbf{x})\, d\mathbf{x} \tag{36}$$

(also called "empirical Bayes" approach [24]) and then plug in this estimate in a MAP or Bayesian criterion, as in the standard (nonblind) restoration problem:

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} p(\mathbf{y}|\mathbf{x}, \hat{\theta}_{ML}) p(\mathbf{x}). \tag{37}$$

Sometimes the plug-in estimate is not $\hat{\theta}_{ML}$, but some other estimate. Under regularity conditions, one may expect $\hat{\theta}_{ML}$ to be an accurate estimator of θ. In this case, the plug-in method works well.

2. Joint estimation of θ and $\mathbf{x}$:

$$(\hat{\mathbf{x}}, \hat{\theta}) = \arg\max_{\mathbf{x},\theta} p(\mathbf{y}|\mathbf{x}, \theta) p(\mathbf{x}). \tag{38}$$

EM algorithms have successfully been used for solving such problems [29, Ch. 7]. The incomplete data are $\mathbf{y}$, the complete data are $(\mathbf{x}, \mathbf{y})$, and estimates of θ are obtained iteratively. In addition, each E-step yields an updated estimate of $\mathbf{x}$.

3. If a prior $p(\theta)$ is available, the estimation problem may be formulated as a MAP estimation problem:

$$(\hat{\mathbf{x}}_{MAP}, \hat{\theta}_{MAP}) = \arg\max_{\mathbf{x},\theta} p(\mathbf{y}|\mathbf{x}, \theta) p(\mathbf{x}) p(\theta) \tag{39}$$

where we have assumed that θ and $\mathbf{X}$ are independent. The problem can again be solved using an EM algorithm.

In all these problems, the size of the parameter vector θ could be large. In the linear blur problem for instance, the impulse response h of the blur filter might be completely unknown and viewed as θ.

You and Kaveh [30] have proposed an alternating maximization algorithm in which the estimate $\hat{h}^{(k)}$ of the impulse response at step k of the algorithm is fixed, and the MAP criterion $\mathcal{E}(\mathbf{x}, \hat{h}^{(k)})$ is maximized over $\mathbf{x}$ (standard nonblind restoration problem), producing an estimate $\hat{\mathbf{x}}^{(k)}$. Then $\hat{\mathbf{x}}^{(k)}$ is held fixed, and the MAP criterion is maximized over h to yield $\hat{h}^{(k+1)}$, etc. This algorithm is greedy, and the sequence $\mathcal{E}(\hat{\mathbf{x}}^{(k)}, \hat{h}^{(k)})$ is nondecreasing in k.

12.4 Restoration Using Hyperpriors

Often a good prior model for the image $\mathbf{x}$ may be available, but some parameters of that model may be unknown. For instance, the wavelet coefficients of $\mathbf{x}$ may be modeled as independent Laplacian random variables with subband-dependent variances, but these variances may be unknown. Let θ be the vector of unknown parameters.

Then the prior can be written as $p(\mathbf{x}|\theta)$, and the estimation problem is analogous to the blind deconvolution problem (with unknown degradation model parameters) discussed above. In particular, one may assume a prior $p(\theta)$ on θ, in which case

$$p(\mathbf{x}) = \int p(\mathbf{x}|\theta) p(\theta)\, d\theta.$$

Here θ is called *hyperparameter*, and $p(\theta)$ is called *hyperprior*. This is also known as a hierarchical prior model [11].

Also note that due to the equivalence between MAP estimation and deterministic regularization, the problem of choosing the regularization parameter λ is equivalent to the problem of choosing the scale parameter λ in the equivalent prior

$$p(\mathbf{x}|\lambda) = \frac{1}{Z(\lambda)} \exp\left\{-\lambda\phi(\mathbf{x})\right\}.$$

13 Exponential Families

We have observed that statistical inference for Gaussian and Poisson processes is often tractable. Moreover, the calculations involved present clear similarities. This is due to the fact that Gaussian and Poisson distributions possess a common mathematical structure – they are elements of so-called exponential families of distributions. This section covers the basic theory of exponential families and suggests convenient ways to choose prior distributions on the parameters; these distributions are so-called conjugate priors. An excellent reference for this material is Lehmann and Casella's textbook [31].

Let $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_d)$ be a d-dimensional parameter taking values in a parameter space Θ, $h(x)$ a real-valued function of x, and $\mathbf{u}(x) = (u_1(x), \cdots, u_d(x))$ be a collection of d real-valued functions of x. The family of pdf's indexed by $\theta \in \Theta$,

$$p_{\boldsymbol{\theta}}(x) = h(x) \exp\{\boldsymbol{\theta}^T \mathbf{u}(x) - A(\boldsymbol{\theta})\}, \tag{40}$$

is said to be a *canonical d-dimensional exponential family*. The normalization function $A(\boldsymbol{\theta})$ that appears in (40) is given by

$$A(\boldsymbol{\theta}) = \ln \int h(x) \exp\{\boldsymbol{\theta}^T \mathbf{u}(x)\}\, dx$$

so that $\int p_{\boldsymbol{\theta}} = 1$ for all $\boldsymbol{\theta} \in \Theta$.

Exponential families are also encountered in *noncanonical form*,

$$p_{\boldsymbol{\theta}}(x) = h(x) \exp\{\boldsymbol{\eta}^T(\boldsymbol{\theta}) \mathbf{u}(x) - B(\boldsymbol{\theta})\},$$

which can be reduced to the canonical form (40) by using $\boldsymbol{\eta}$ as the parameter of the family.

13.1 Examples

Poisson distribution:

$$p_\lambda(x) = \frac{\lambda^x}{x!} e^{-\lambda} \ = \frac{1}{x!} e^{x \ln \lambda - \lambda}$$

which can be viewed a canonical 1-D exponential family

$$p_\theta(x) = \frac{1}{x!} e^{\theta x - e^\theta}$$

where $\Theta = \mathbb{R}$ and

$$\theta = \ln \lambda, \quad u(x) = x, \quad A(\theta) = e^\theta, \quad h(x) = \frac{1}{x!}.$$

Gaussian distribution:

$$\begin{aligned} p_{\mu,\sigma^2}(x) &= (2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \\ &= (2\pi)^{-1/2} \exp\left\{ \frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x^2 - \frac{\mu^2}{2\sigma^2} + \frac{1}{2} \ln \frac{1}{\sigma^2} \right\} \end{aligned}$$

which can be viewed as a canonical 2-D exponential family

$$p_{\boldsymbol{\theta}}(x) = (2\pi)^{-1/2} \exp\left\{ \theta_1 x + \theta_2 x^2 + \frac{\theta_1^2}{4\theta_2} + \frac{1}{2} \ln(-2\theta_2) \right\}$$

where $\Theta = \mathbb{R} \times \mathbb{R}^-$ and

$$\begin{aligned} \theta_1 = \frac{\mu}{\sigma^2} \quad &, \quad u_1(x) = x, \\ \theta_2 = -\frac{1}{2\sigma^2} \quad &, \quad u_2(x) = x^2, \\ A(\boldsymbol{\theta}) = \frac{\theta_1^2}{-4\theta_2} - \frac{1}{2} \ln(-2\theta_2) \quad &, \quad h(x) = (2\pi)^{-1/2}. \end{aligned}$$

13.2 Properties

Moments. Differentiating both sides of the identity

$$1 = \int h(x) e^{\boldsymbol{\theta}^T \mathbf{u}(x) - A(\boldsymbol{\theta})} \, dx$$

with respect to $\boldsymbol{\theta}$, we obtain the vector equality

$$\begin{aligned} \mathbf{0} &= \nabla_{\boldsymbol{\theta}} \int h(x) e^{\boldsymbol{\theta}^T \mathbf{u}(x) - A(\boldsymbol{\theta})} \, dx \\ &= \int h(x) e^{\boldsymbol{\theta}^T \mathbf{u}(x) - A(\boldsymbol{\theta})} [\mathbf{u}(x) - \nabla A(\boldsymbol{\theta})] \, dx \\ &= \mathbb{E}[\mathbf{u}(X)] - \nabla A(\boldsymbol{\theta}). \end{aligned}$$

Differentiating a second time, we obtain

$$\begin{aligned}
\mathbf{0} &= \nabla_{\boldsymbol{\theta}}^2 \int h(x) e^{\boldsymbol{\theta}^T \mathbf{u}(x) - A(\boldsymbol{\theta})}\, dx \\
&= \int h(x) e^{\boldsymbol{\theta}^T \mathbf{u}(x) - A(\boldsymbol{\theta})} \left([\mathbf{u}(x) - \nabla A(\boldsymbol{\theta})][\mathbf{u}(x) - \nabla A(\boldsymbol{\theta})]^T - \nabla^2 A(\boldsymbol{\theta}) \right) dx \\
&= \mathbb{E}\left([\mathbf{u}(X) - \nabla A(\boldsymbol{\theta})][\mathbf{u}(X) - \nabla A(\boldsymbol{\theta})]^T \right) - \nabla^2 A(\boldsymbol{\theta}) \\
&= \mathrm{Cov}[\mathbf{u}(X)\mathbf{u}^T(X)] - \nabla^2 A(\boldsymbol{\theta}).
\end{aligned}$$

Hence the mean and covariance of $\mathbf{u}(X)$ are given by

$$\mathbb{E}[\mathbf{u}(X)] = \nabla A(\boldsymbol{\theta}) \tag{41}$$

$$\mathrm{Cov}[\mathbf{u}(X)\mathbf{u}^T(X)] = \nabla^2 A(\boldsymbol{\theta}). \tag{42}$$

For the Poisson example ($d = 1$) of Sec. 13.1, we have

$$A(\theta) = \nabla A(\theta) = \nabla^2 A(\theta) = e^{\theta}.$$

For the Gaussian example ($d = 2$), we have

$$A(\boldsymbol{\theta}) = -\frac{\theta_1^2}{4\theta_2} - \frac{1}{2}\ln(-2\theta_2), \quad \nabla A(\boldsymbol{\theta}) = \begin{pmatrix} -\frac{\theta_1}{2\theta_2} \\ \frac{\theta_1^2}{4\theta_2^2} - \frac{1}{2\theta_2} \end{pmatrix}, \quad \nabla^2 A(\boldsymbol{\theta}) = \begin{pmatrix} -\frac{1}{2\theta_2} & \frac{\theta_1}{2\theta_2^2} \\ \frac{\theta_1}{2\theta_2^2} & -\frac{\theta_1^2}{2\theta_2^3} + \frac{1}{2\theta_2^2} \end{pmatrix}.$$

It will be convenient to view

$$\boldsymbol{\xi} = H(\boldsymbol{\theta}) \triangleq \nabla A(\boldsymbol{\theta}) = \mathbb{E}[\mathbf{u}(X)] \tag{43}$$

as a transformation of the vector $\boldsymbol{\theta}$. In view of the last equality in (43), $\boldsymbol{\xi}$ is also called the *mean-value parameterization* of the exponential family [31, p. 126]. We will assume that the Jacobian of $H = \nabla a$,

$$\mathbf{J} \triangleq \left\{ \frac{\partial H_i}{\partial \theta_j} \right\}_{i,j=1}^{d} = \nabla^2 A(\boldsymbol{\theta}) = \mathrm{Cov}[\mathbf{u}(X)\mathbf{u}^T(X)],$$

has full rank and is therefore invertible.

Fisher Information. From (40), the gradient of the loglikelihood function is given by

$$\nabla \ln p_{\boldsymbol{\theta}}(x) = \mathbf{u}(x) - \nabla A(\boldsymbol{\theta}).$$

Thus Fisher information is obtained as

$$\begin{aligned}
\mathbf{I}_{\boldsymbol{\theta}} &= \mathbb{E}\left\{ (\nabla \ln p_{\boldsymbol{\theta}}(X))(\nabla \ln p_{\boldsymbol{\theta}}(X))^T \right\} \\
&= \mathbb{E}\left\{ [\mathbf{u}(X) - \nabla A(\boldsymbol{\theta})][\mathbf{u}(X) - \nabla A(\boldsymbol{\theta})]^T \right\} \\
&= \mathrm{Cov}[\mathbf{u}(X)\mathbf{u}^T(X)] \qquad (44) \\
&= \nabla^2 A(\boldsymbol{\theta}). \qquad (45)
\end{aligned}$$

ML Estimation. The likelihood equation is given by

$$\begin{aligned} 0 &= \nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(x)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{ML}} \\ &= \mathbf{u}(x) - \nabla A(\hat{\boldsymbol{\theta}}_{ML}). \end{aligned}$$

Since ML estimators commute with nonlinear operations, the ML estimate of the transformed vector $\boldsymbol{\xi}$ defined in (43) is simply

$$\hat{\boldsymbol{\xi}}_{ML} = \nabla A(\hat{\boldsymbol{\theta}}_{ML}) = \mathbf{u}(x).$$

Identifiability. If there exist $\boldsymbol{\theta}' \neq \boldsymbol{\theta}$ such that the distributions $P_{\boldsymbol{\theta}'}$ and $P_{\boldsymbol{\theta}}$ are identical, then $\boldsymbol{\theta}$ is said to be *unidentifiable*.

I.i.d. Data. Let $\mathbf{x} = (x_1, \cdots, x_N)$ be an iid sequence drawn from the exponential distribution (40). Then

$$\begin{aligned} p_{\boldsymbol{\theta}}(\mathbf{x}) &= \prod_{i=1}^{N} p_{\boldsymbol{\theta}}(x_i) \\ &= \left(\prod_{i=1}^{N} h(x_i)\right) \exp\left\{\boldsymbol{\theta}^T \sum_{i=1}^{N} \mathbf{u}(x_i) + NA(\boldsymbol{\theta})\right\}, \quad \boldsymbol{\theta} \in \Theta, \end{aligned}$$

is still an exponential family. The ML estimate of $\boldsymbol{\theta}$ given $\mathbf{X}$ is given by

$$\hat{\boldsymbol{\xi}}_{ML} = \nabla A(\hat{\boldsymbol{\theta}}_{ML}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{u}(x_i). \tag{46}$$

Asymptotics. The convergence of $\hat{\boldsymbol{\theta}}_{ML}$ to the true $\boldsymbol{\theta}$ can be studied from (46). Recall from (44) that the covariance matrix for $\mathbf{u}(X)$ is the Fisher information matrix $\mathbf{I}_\theta$. By the Central Limit Theorem, the asymptotic distribution of the the sample average $\frac{1}{N}\sum_{i=1}^{N} \mathbf{u}(X_i)$ is Gaussian, more specifically

$$\begin{aligned} \sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{u}(X_i) - \mathbb{E}[\mathbf{u}(X)]\right) &\xrightarrow{d} \mathcal{N}(0, \mathbf{I}_{\boldsymbol{\theta}}) \\ \sqrt{N}\left(\hat{\boldsymbol{\xi}}_{ML} - \boldsymbol{\xi}\right) &\xrightarrow{d} \mathcal{N}(0, \mathbf{I}_{\boldsymbol{\theta}}) \\ \sqrt{N}\left(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}\right) &\xrightarrow{d} \mathcal{N}(0, \mathbf{I}_{\boldsymbol{\theta}}^{-1}) \end{aligned} \tag{47}$$

where the second line follows from (43) and (46), and the last line follows from our assumption that the Jacobian $\mathbf{J}$ of the transformation $\boldsymbol{\xi} = \nabla A(\boldsymbol{\theta})$ is invertible and the fact that $\mathbf{J}^{-1}\mathbf{I}_{\boldsymbol{\theta}}\mathbf{J} = \mathbf{I}_{\boldsymbol{\theta}}^{-1}$. From (47) we conclude that the ML estimator $\hat{\boldsymbol{\theta}}_{ML}$ is asymptotically efficient [31, p. 439], i.e., it asymptotically satisfies the Cramer-Rao bound for unbiased estimators.

13.3 Conjugate Priors

For a given distribution $p(x|\boldsymbol{\theta})$, there often exists several prior distributions $p(\boldsymbol{\theta})$ that are "well matched" to $p(x|\boldsymbol{\theta})$ in the sense that the posterior distribution $p(\boldsymbol{\theta}|x)$ is tractable. For instance, if both $p(x|\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ are Gaussian, then so is $p(\boldsymbol{\theta}|x)$.

More generally, a similar concept holds when $p(x|\boldsymbol{\theta})$ is taken from an arbitrary exponential family. Consider the family (40) and the following *conjugate prior*, which is parameterized by a number $\nu \in \mathbb{R}$ and a vector $\mathbf{v} \in \mathbb{R}^d$:

$$p(\boldsymbol{\theta}) = \tilde{h}(\nu, \mathbf{v}) \exp\{\nu \boldsymbol{\theta}^T \mathbf{v} - \nu A(\boldsymbol{\theta})\}. \tag{48}$$

In the absence of a prior, we have seen in (46) that ML estimation of $\boldsymbol{\theta}$ from iid samples is easy. Consider again $\mathbf{X} = (X_1, \cdots, X_N)$, drawn iid from (40). Then

$$p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = \left(\prod_{i=1}^N h(x_i)\right) \tilde{h}(\nu, \mathbf{v}) \exp\left\{\boldsymbol{\theta}^T \left(\sum_{i=1}^N \mathbf{u}(x_i) + \nu \mathbf{v}\right) - (N+\nu) A(\boldsymbol{\theta})\right\}.$$

Therefore

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto \exp\left\{\boldsymbol{\theta}^T \left(\sum_{i=1}^N \mathbf{u}(x_i) + \nu \mathbf{v}\right) - (N+\nu) A(\boldsymbol{\theta})\right\}$$

has the same form as the prior (48), with

$$\nu' = N + \nu \quad \text{and} \quad \mathbf{v}' = \frac{1}{N+\nu}\left(\sum_{i=1}^N \mathbf{u}(x_i) + \nu \mathbf{v}\right).$$

The MAP estimator is the solution to

$$0 = \nabla \ln p(\boldsymbol{\theta}|\mathbf{x}) \,|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{MAP}},$$

hence

$$\hat{\boldsymbol{\xi}}_{MAP} = \nabla A(\hat{\boldsymbol{\theta}}_{MAP}) = \frac{1}{N+\nu}\left(\sum_{i=1}^N \mathbf{u}(x_i) + \nu \mathbf{v}\right).$$

Observe the remarkable similarity of this expression to (46). It is convenient to think of ν as a *number of pseudo-observations*, and of $\mathbf{v}$ as a prior mean for $\mathbf{u}(X)$. The total Fisher information matrix is

$$\mathbf{I}_{\boldsymbol{\theta}} = (N+\nu)\nabla^2 A(\boldsymbol{\theta}).$$

Example. Consider $\mathbf{X} = (X_1, \cdots, X_N)$ drawn iid from $\mathcal{N}(0, \sigma^2)$. Let $\theta = \frac{1}{2\sigma^2}$, then

$$p(\mathbf{x}|\theta) = \pi^{-N/2} \theta^{N/2} \exp\left\{-\theta \sum_{i=1}^N x_i^2\right\}.$$

Consider the $\Gamma(a, b)$ prior for θ:

$$p(\theta) = \frac{1}{\Gamma(a)b^a}\theta^{a-1}e^{-\theta/b}, \quad a, b, \theta > 0.$$

Hence

$$p(\theta|\mathbf{x}) \propto \theta^{N/2+a-1} \exp\left\{-\theta\left(\frac{1}{b} + \sum_{i=1}^{N} x_i^2\right)\right\}$$

is a $\Gamma(a', b')$ distribution with

$$a' = \frac{N}{2} + a, \quad \text{and} \quad \frac{1}{b'} = \frac{1}{b} + \sum_{i=1}^{N} x_i^2.$$

The MAP estimator of θ given $\mathbf{x}$ is equal to

$$\hat{\theta}_{MAP} = b'(a' - 1) = \frac{\frac{N}{2} + a - 1}{\frac{1}{b} + \sum_{i=1}^{N} x_i^2}$$

and the total Fisher information by

$$I_\theta = \frac{N + 2(a-1)}{2\theta^2}.$$

(Prove this as an exercise.)

Conditional Expectation. We have the following property [31, p. 286]:

$$\mathbb{E}[\boldsymbol{\xi}] = \mathbb{E}[\nabla A(\boldsymbol{\theta})] = \mathbf{v} \tag{49}$$

where the expectation is taken with respect to the prior $p(\boldsymbol{\theta})$. Similarly, for the problem with iid data $\mathbf{x} = (X_1, \cdots, X_N)$, we obtain

$$\mathbb{E}[\boldsymbol{\xi}|\mathbf{x}] = \mathbb{E}[\nabla A(\boldsymbol{\theta})|\mathbf{x}] = \frac{1}{N+\nu}\left(\sum_{i=1}^{N} \mathbf{u}(x_i) + \nu\mathbf{v}\right) \tag{50}$$

14 Maximum-Entropy Priors

The notion of maximum-entropy (maxent) distributions first appeared in connection with inference problems in statistical physics, see Jaynes [32, 33]. The basic principle is that particles in a system tend to be distributed in a way that maximizes randomness, subject to external constraints such as temperature, pressure, etc. The maxent principle has found numerous applications to signal and image processing as well [34, 35, 36]. It provides a systematic, elegant way to construct distributions on high-dimensional objects under constraints on moments of these distributions.

The basic problem is as follows. Find the pdf $p(\mathbf{x})$ that maximizes the differential entropy

$$h(p) \triangleq -\int p(\mathbf{x}) \ln p(\mathbf{x})\, d\mathbf{x} \tag{51}$$

subject to the moment constraints

$$\int p(\mathbf{x}) u_i(\mathbf{x})\, d\mathbf{x} = \mu_i, \quad i = 1, \cdots, d+1. \tag{52}$$

In (52), we fix $u_{d+1}(\mathbf{x}) \equiv 1$ and $\mu_{d+1} = 1$ to account for the pdf constraint $\int p(\mathbf{x})\, d\mathbf{x} = 1$.

14.1 Solution to Maxent Problem

The maximization problem above is concave and is solved using a Lagrangian method. Introducing Lagrange multipliers $\lambda_1, \cdots, \lambda_{d+1}$ corresponding to the moment constraints (52), we define the Lagrangian

$$L(p, \boldsymbol{\lambda}) \;=\; -\int p(\mathbf{x}) \ln p(\mathbf{x})\, d\mathbf{x} + \sum_{i=1}^{d+1} \lambda_i \left(\int p(\mathbf{x}) u_i(\mathbf{x})\, d\mathbf{x} - \mu_i \right) \tag{53}$$

Setting partial derivatives with respect to each $p(\mathbf{x})$ to zero, we obtain

$$0 = \frac{\partial L(p, \boldsymbol{\lambda})}{\partial p(\mathbf{x})} = -\ln p(\mathbf{x}) + \sum_{i=1}^{d} \lambda_i u_i(\mathbf{x}) + \lambda_{d+1} - 1, \quad \forall \mathbf{x}.$$

Hence the maxent solution, if it exists, is of the exponential form (recall (40))

$$\begin{aligned} p(\mathbf{x}) &= \exp\left\{ \sum_{i=1}^{d} \lambda_i u_i(\mathbf{x}) + \lambda_{d+1} - 1 \right\} \\ &= \frac{1}{Z(\boldsymbol{\lambda})} \exp\left\{ \sum_{i=1}^{d} \lambda_i u_i(\mathbf{x}) \right\} \end{aligned} \tag{54}$$

where we let $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_d)$.

The condition for existence of the maxent solution is as follows. [2]

Theorem. Assume p is of the exponential form (54) and satisfies the moment constraints (52). Then p is the unique maxent solution.

14.2 Model Fitting

Assume that we are trying to model a pdf $p(\mathbf{x})$ and that the only information available to us is in the form of the moment constraints (52). From the discussion above we know that the maxent solution $\hat{p}_d(\mathbf{x})$ is of exponential form:

$$\hat{p}_d(\mathbf{x}) = \frac{1}{Z(\boldsymbol{\lambda})} \exp\left\{\sum_{i=1}^{d} \lambda_i u_i(\mathbf{x})\right\}.$$

We now consider two questions:

1. How well does the maxent solution $\hat{p}_d$ approximate the true distribution p?
2. What are the benefits of including more constraints into the model?

First observe that adding constraints cannot increase the value of the maximum, so the sequence $h(\hat{p}_d)$ is nonincreasing:

$$h(\hat{p}_1) \geq h(\hat{p}_2) \geq \cdots \geq h(p).$$

Second, if we measure the goodness of fit of $\hat{p}_d$ to p using Kullback-Leibler divergence:

$$D(p\|\hat{p}_d) = \int p(\mathbf{x}) \ln \frac{p(\mathbf{x})}{\hat{p}_d(\mathbf{x})}\, d\mathbf{x},$$

we have the following interesting result due to Csiszár [39] and rediscovered by Zhu *et al.* [34] in a computer vision context.

Theorem If $\hat{p}_d$ is the maxent prior, then

$$D(p\|\hat{p}_d) = h(\hat{p}_d) - h(p). \tag{55}$$

[2] Csiszár [37] gives a simple example in which the maxent solution does not exist: let the constraints be $\mathbb{E}(X) = 0$, $\mathbb{E}(X^2) = 1$, and $\mathbb{E}(X^3) = 1$. To have $Z(\boldsymbol{\lambda}) < \infty$ in (54) we need $\lambda_3 = 0$, in which case (54) is Gaussian and therefore $\mathbb{E}(X^3) = 0$, which violates the constraint $\mathbb{E}(X^3) = 1$.

Proof. Denoting by $\mathbb{E}_p[v(\mathbf{X})] = \int p(\mathbf{x})v(\mathbf{x})\,d\mathbf{x}$ the expectation of a function $v(\mathbf{X})$ under probability law p, we write

$$\begin{aligned} D(p\|\hat{p}_d) &= \int p(\mathbf{x}) \ln p(\mathbf{x})\,d\mathbf{x} - \int p(\mathbf{x}) \ln \hat{p}_d(\mathbf{x})\,d\mathbf{x} \\ &= -h(p) - \mathbb{E}_p[\ln \hat{p}_d(\mathbf{X})] \\ &\overset{(a)}{=} -h(p) + \ln Z(\boldsymbol{\lambda}) - \boldsymbol{\lambda}^T \mathbb{E}_p[\mathbf{u}(\mathbf{X})] \\ &\overset{(b)}{=} -h(p) + \ln Z(\boldsymbol{\lambda}) - \boldsymbol{\lambda}^T \boldsymbol{\mu} \\ &\overset{(c)}{=} -h(p) + \ln Z(\boldsymbol{\lambda}) - \boldsymbol{\lambda}^T \mathbb{E}_{\hat{p}_d}[\mathbf{u}(\mathbf{X})] \\ &= -h(p) - \mathbb{E}_{\hat{p}_d}[\ln \hat{p}_d(\mathbf{X})] \\ &= -h(p) + h(\hat{p}_d) \end{aligned}$$

where equality (a) is due to the exponential form of $\hat{p}_d$, and (b) and (c) are due to the fact that both p and $\hat{p}_d$ satisfy the moment constraints (52). □

14.3 Feature Pursuit

The above theorem cannot be used to quantify the goodness of the fit because $h(p)$ in the right side of (55) is unknown. However the theorem suggests that selecting moment constraints that lead to the greatest reduction in the maxent entropy $h(\hat{p}_d)$ is beneficial. The papers by Zhu *et al.* [34, 35] explore this idea in the context of image texture modeling. They define a large set $\mathcal{F}$ of feature mappings $u_i(\cdot)$, including energy and coarse histograms of filtered images. The problem of selecting d best features out of $|\mathcal{F}|$ is of combinatorial complexity, hence they explore a greedy strategy in which features are added one at a time to the model. Models of increasing order are thus nested. Zhu *et al.* have termed his greedy strategy *feature pursuit*.

Applications of feature pursuit to texture modeling have been promising: the resulting maxent models are capable of generating interesting textures such as mud ground, cheetah skin, fabrics and textons, with a high degree of realism.

15 Alternating Minimization Algorithms

The EM algorithm may be formulated as an alternating minimization algorithm, as studied by Csiszár and Tusnady [39]. This interpretation leads to a family of variants of the EM algorithm that in some cases are easier to implement and have good convergence properties.

15.1 EM Algorithm Revisited

Assume for simplicity that the random variable Z in the complete-data space is discrete. The incomplete data are given by $Y = h(Z)$ where h is a many-to-one mapping. The problem is to maximize the incomplete-data loglikelihood and evaluate

$$\hat{\theta}_{ML} = \arg\max_{\theta\in\Theta} \ln p(y|\theta).$$

The pmf for the complete data is denoted by $p(z|\theta)$ and is assumed to be strictly positive for all z in the complete data space and for all $\theta \in \Theta$. The pmf for the incomplete data is the marginal

$$p(y|\theta) = \sum_{z\in h^{-1}(y)} p(z|\theta),$$

where $h^{-1}(y)$ denotes the set of z such that $h(z) = y$.

For fixed y, we use the shorthand $q_\theta(z) = p(z|y,\theta)$, a pmf whose support set is $h^{-1}(y)$. We also denote by $q(z)$ a surrogate pmf over the complete data space. The entropy of a pmf $q(z)$ is defined as

$$H(p) = -\sum_z q(z) \ln q(z) \geq 0. \tag{56}$$

The Kullback-Leibler divergence (or relative entropy) between two pmf's $q(z)$ and $q'(z)$ is defined as

$$D(q\|q') \triangleq \sum_z q(z) \ln \frac{q(z)}{q'(z)} \geq 0, \tag{57}$$

with equality if and only if $q = q'$. (The natural logarithm is used for convenience.) We also use the notational convention $0 \ln 0 = 0$.

Recall the EM algorithm first initializes $\hat{\theta}^{(0)}$. At iteration $k+1$, the current estimate $\hat{\theta}^{(k)}$ is updated as follows:

E-Step: evaluate $Q(\theta|\hat{\theta}^{(k)}) \triangleq \sum_z p(z|y,\hat{\theta}^k) \ln p(z|\theta)$;

M-Step: let $\hat{\theta}^{(k+1)} = \arg\max_\theta Q(\theta|\hat{\theta}^{(k)})$.

Let us now derive an alternative interpretation of the E- and M-steps. For any pmf $q(z)$ defined over the complete data space, we may write

$$\begin{aligned}
\ln p(y|\theta) &= \sum_z q(z) \ln p(y|\theta) \\
&= \sum_z q(z) \ln \frac{p(y|\theta) q_\theta(z)}{q(z)} + \sum_z q(z) \ln \frac{q(z)}{q_\theta(z)} \\
&= \sum_z q(z) \ln \frac{p(y,z|\theta)}{q(z)} + \sum_z q(z) \ln \frac{q(z)}{q_\theta(z)} \\
&= L(q,\theta) + D(q\|q_\theta)
\end{aligned}$$

where we have defined the function

$$L(q,\theta) \triangleq \sum_z q(z) \ln \frac{p(y,z|\theta)}{q(z)}. \tag{58}$$

Since $D(q\|q_\theta) \geq 0$, we have

$$L(q,\theta) \leq \ln p(y|\theta),$$

i.e, $L(q,\theta)$ is a lower bound on the loglikelihood, for any choice of q and θ. Moreover

$$\begin{aligned}
\arg\max_q L(q,\theta) &= \arg\max_q [\ln p(y|\theta) - D(q\|q_\theta)] = \arg\min_q D(q\|q_\theta) = q_\theta \\
\max_q L(q,\theta) &= \ln p(y|\theta)
\end{aligned}$$

for any value of θ.

Define the shorthand

$$q^{(k)}(z) = q_{\hat{\theta}^{(k)}}(z) = p(z|y,\hat{\theta}^{(k)}), \tag{59}$$

and recall that the support set of the pmf $q^{(k)}$ is $h^{-1}(y)$. Write

$$\begin{aligned}
Q(\theta|\hat{\theta}^{(k)}) &= \sum_z q^{(k)}(z) \ln p(z|\theta) \\
&= \sum_z q^{(k)}(z) \ln p(y,z|\theta) \\
&= \sum_z q^{(k)}(z) \ln \frac{p(y,z|\theta)}{q^{(k)}(z)} + \sum_z q^{(k)}(z) \ln q^{(k)}(z) \\
&= L(q^{(k)},\theta) - H(q^{(k)})
\end{aligned}$$

where the second equality holds because the only nonzero terms in the sum are those indexed by $z \in h^{-1}(y)$, in which case $p(y|z,\theta) = 1$ and therefore $p(z|\theta) = p(y,z|\theta)$. From the last line we conclude that

$$\arg\max_\theta Q(\theta|\hat{\theta}^{(k)}) = \arg\max_\theta L(q^{(k)},\theta).$$

Initialize $\hat{\theta}^{(0)}$.
Then for $k = 0, 1, \cdots$, perform the following operations:
$q^{(k)} = \arg\max_q L(q, \hat{\theta}^{(k)}) \quad \Longrightarrow \quad L(q^{(k)}, \hat{\theta}^{(k)}) = \ln p(y\|\hat{\theta}^{(k)})$;
$\hat{\theta}^{(k+1)} = \arg\max_\theta L(q^{(k)}, \theta)$.

Table 1: Alternating maximization of $L(q, \theta)$.

.

Hence the EM algorithm is equivalent to the *alternating maximization algorithm* of Table 1, in which the function $L(q, \theta)$ is maximized over q with θ fixed, then over θ with q fixed, and this two-step process is repeated until convergence.

Therefore the sequence of loglikelihoods $\ln p(y|\hat{\theta}^{(k)})$ is nondecreasing, and convergence of the EM algorithm to a stable value of the loglikelihood function is guaranteed.

15.2 Alternating Minimization of Kullback-Leibler Divergence

The alternating maximization procedure admits another interesting interpretation. Recall that $p(y|z, \theta) = 1$ for all $z \in h^{-1}(y)$ and thus

$$p(y, z|\theta) = p(z|\theta)p(y|z, \theta) = p(z|\theta), \quad \forall z \in h^{-1}(y). \tag{60}$$

Denote by $\mathcal{Q}$ the set of pmf's $q(z)$ whose support set is $h^{-1}(y)$. From (60) and (58), we can write

$$L(q, \theta) = -\sum_z q(z) \ln \frac{q(z)}{p(z|\theta)} = -D(q\|p(\cdot|\theta)), \quad \forall q \in \mathcal{Q}.$$

Using the shorthand

$$r^{(k)}(z) = p(z|\hat{\theta}^{(k)})$$

and observing that $q^{(k)}$ defined in (59) belongs to $\mathcal{Q}$, we see that the algorithm of Table 1 is equivalent to the algorithm of Table 2 for minimizing Kullback-Leibler divergence. Here the *model family* $\mathcal{R}$ is the set of pmf's $p(z|\theta)$ parameterized by $\theta \in \Theta$. This iterative minimization procedure is illustrated in Fig. 15.

Initialize $r^{(0)} \in \mathcal{R}$.
Then for $k = 0, 1, \cdots$, perform the following operations:
$q^{(k)} = \arg\min_{q \in \mathcal{Q}} D(q\|r^{(k)})$
$r^{(k+1)} = \arg\min_{r \in \mathcal{R}} D(q^{(k)}\|r)$.

Table 2: Alternating minimization of Kullback-Leibler divergence

.

The Kullback-Leibler divergence $D(q\|r)$ is convex in the pair (q, r). As shown by Csiszár and Tusnady [39], if the sets $\mathcal{Q}$ and $\mathcal{R}$ are convex, the algorithm of Table 2 is guaranteed

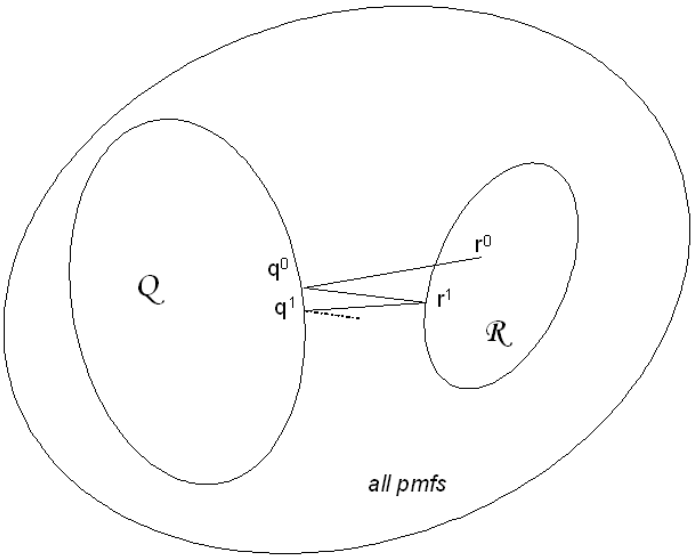


Figure 15: A pictorial view of alternating minimization.

to converge to a global minimum (over $\mathcal{Q} \times \mathcal{R}$) of the function $D(q\|r)$. In our problem, $\mathcal{Q}$ is convex, so a sufficient condition for convergence to a global minimum is that the model family $\mathcal{R}$ be convex. In terms of our original EM problem, this ensures convergence to a global maximum of the loglikelihood function $\ln p(y|\theta)$.

15.3 Extensions

In some cases, finding the maximizing θ in the M-step of the EM algorithm is difficult. In other cases, computing the expectation in the E-step is difficult because the conditional expectation $p(z|y, \hat{\theta}^{(k)})$ does not admit a simple form. Several generalizations of the EM algorithm have been proposed to deal with such difficulties. See Gunawardana and Byrne [40] for an elegant analysis of the convergence properties of such generalizations.

- *Generalized EM*: In the M-step of the EM algorithm, one does not perform a full optimization over θ but instead look for some $\hat{\theta}^{(k+1)}$ that increases the value of the cost function:

  $$Q(\hat{\theta}^{(k+1)}|\hat{\theta}^{(k)}) \geq Q(\hat{\theta}^{(k)}|\hat{\theta}^{(k)})$$

  or equivalently,

  $$L(q^{(k)}, \hat{\theta}^{(k+1)}) \geq L(q^{(k)}, \hat{\theta}^{(k)}).$$

From the algorithm in Table 1, we see that the monotonicity property of the likelihood function is preserved.

- *Variational EM*: Use a tractable surrogate distribution $q^{(k)}$ in place of the ideal $p(z|y, \hat{\theta}^{(k)})$. For instance, choose $q^{(k)}$ from an exponential family, or choose $q^{(k)}$ to be a product distribution. Given a family $\mathcal{Q}$ of tractable candidate distributions, variational EM selects $q^{(k)} \in \mathcal{Q}$ that maximizes $L(\cdot, \hat{\theta}^{(k)})$. Unlike the EM algorithm, monotonicity of the likelihood function is not guaranteed (unless $\mathcal{Q}$ is the class of all pmf's, in which case variational EM trivially reduces to EM).

16 Regularization Methods and PDEs

There exist interesting connections between some of the regularization methods studied earlier in this chapter and partial differential equations (PDEs). The idea is to formulate the regularization problem in terms of images defined over continuous domains and view the regularized solution as the limit of the temporal evolution of a progressively restored image.

16.1 Gradient Descent for Discrete Domains

First we set up a numerical algorithm for computing regularized estimates when the images $\mathbf{x} = \{x(t), t \in \mathcal{T}\}$ and $\mathbf{y} = \{y(u), u \in \mathcal{U}\}$ are defined over discrete domains $\mathcal{T}$ and $\mathcal{U}$, respectively. The gradient of the cost function of (17),

$$\mathcal{E}(\mathbf{x}) = \|\mathbf{y} - H\mathbf{x}\|^2 + \lambda\phi(\mathbf{x}),$$

is given by

$$\nabla_{\mathbf{x}}\mathcal{E}(\mathbf{x}) = 2H^T(H\mathbf{x} - \mathbf{y}) + \lambda\nabla_{\mathbf{x}}\phi(\mathbf{x}).$$

The gradient is equal to zero for $\mathbf{x} = \hat{\mathbf{x}}_{REG}$. Typical gradient-descent algorithms start with an estimate $\hat{\mathbf{x}}^{(k)}$ at step $k = 0$ and update $\hat{\mathbf{x}}^{(k)}$ by moving in the direction of the negative gradient:

$$\hat{\mathbf{x}}^{(k+1)} \quad = \quad \hat{\mathbf{x}}^{(k)} - \epsilon\, \nabla_{\mathbf{x}}\mathcal{E}(\hat{\mathbf{x}}^{(k)}) \tag{61}$$

where ϵ is the step size, which should be neither too small (in which case the algorithm is very slow to converge) nor too large (in which case the algorithm may diverge). If the algorithm converges, we have

$$\lim_{k\to\infty} \hat{\mathbf{x}}^{(k)} = \hat{\mathbf{x}}_{REG}.$$

Local minima of $\mathcal{E}(\mathbf{x})$ are global minima if $\mathcal{E}(\mathbf{x})$ is convex: this is an important computational advantage of convexity. Notice that $\|\mathbf{y} - H\mathbf{x}\|^2$ is a nonnegative definite, quadratic function of $\mathbf{x}$, so $\mathcal{E}(\mathbf{x})$ is guaranteed to be convex if $\phi(\mathbf{x})$ is convex.

The update equation (61) can be written in the form

$$\begin{aligned}\frac{\hat{\mathbf{x}}^{(k+1)} - \hat{\mathbf{x}}^{(k)}}{\epsilon} &= -\nabla_{\mathbf{x}} \mathcal{E}(\hat{\mathbf{x}}^{(k)}) \\ &= 2H^T(\mathbf{y} - H\hat{\mathbf{x}}^{(k)}) - \lambda \nabla_{\mathbf{x}} \phi(\hat{\mathbf{x}}^{(k)}),\end{aligned} \tag{62}$$

a fact which will be useful in the next section.

16.2 Connection to PDEs

Assume now that the images $\mathbf{x} = \{x(t), t \in \mathcal{T}\}$ and $\mathbf{y} = \{y(u), u \in \mathcal{U}\}$ are defined over continuous domains $\mathcal{T}$ and $\mathcal{U}$, respectively. Denote by

$$\nabla_t x(t) = \left(\frac{\partial x(t)}{\partial t_1}, \frac{\partial x(t)}{\partial t_2} \right), \quad t \in \mathcal{T},$$

the spatial gradient of $\mathbf{x}$, and

$$\|\nabla_t x(t)\| = \left[\left(\frac{\partial x(t)}{\partial t_1} \right)^2 + \left(\frac{\partial x(t)}{\partial t_2} \right)^2 \right]^{1/2}.$$

its Euclidean norm. Assume $\phi(\mathbf{x})$ is of the form

$$\phi(\mathbf{x}) = \int_{\mathcal{T}} g(\|\nabla_t x(t)\|)\, dt$$

where $g : \mathbb{R}^+ \to \mathbb{R}^+$ is strictly increasing and twice differentiable. For Tikhonov regularization (Sec. 8) and TV regularization (Sec. 10.1), we have $g(u) = \frac{1}{2}u^2$ and $g(u) = u$, respectively.

Note that $\|\nabla_t x(t)\|$ is a convex function of $\mathbf{x}$ because the triangle inequality implies that

$$\|\nabla_t (a x_1(t) + (1-a)x_2(t))\| \leq a\|\nabla_t x_1(t)\| + (1-a)\|\nabla_t x_2(t)\|$$

for all $a \in [0,1]$ and all signals $\mathbf{x}_1$ and $\mathbf{x}_2$. Hence if $g(\cdot)$ is convex, $\phi(\mathbf{x})$ is also convex.

The partial (Fréchet) derivative of $\phi(\mathbf{x})$ with respect to $x(t)$ is given by

$$\begin{aligned}\frac{\partial}{\partial x(t)} \phi(\mathbf{x}) &= -\mathrm{div} \left(\frac{g'(\|\nabla_t x(t)\|)}{\|\nabla_t x(t)\|} \nabla_t x(t) \right) \\ &= -\mathrm{div} \left(c(\|\nabla_t x(t)\|) \nabla_t x(t) \right), \quad \forall t \in \mathcal{T},\end{aligned} \tag{63}$$

where

$$\mathrm{div} \begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix} = \frac{\partial u_1(t)}{\partial t_1} + \frac{\partial u_2(t)}{\partial t_2}$$

is the divergence operator, and

$$c(u) \triangleq \frac{g'(u)}{u}$$

is the so-called *diffusion coefficient*. It is equal to 1 if ϕ is the Tikhonov regularization functional and to $\frac{1}{u}$ if ϕ is the TV regularization functional.

The divergence term in (63) may be expanded in the form [41]

$$\operatorname{div}\left(c(\|\nabla_t\, x(t)\|)\nabla_t\, x(t)\right) = c(\|\nabla_t\, x(t)\|)\,\Delta^2 x(t) + \nabla_t c(\|\nabla_t\, x(t)\|)\cdot\nabla_t\, x(t) \tag{64}$$

where $\Delta^2 x(t) = \frac{\partial^2 x(t)}{\partial t_1^2} + \frac{\partial^2 x(t)}{\partial t_2^2}$ is the *Laplacian* of x. In the special case of Tikhonov regularization, $c = 1$ and the divergence term simplifies into

$$\operatorname{div}\left(c(\|\nabla_t\, x(t)\|)\,\nabla_t\, x(t)\right) = \Delta^2 x(t). \tag{65}$$

Let us consider the data-fidelity term now. For simplicity, we assume that H is a linear operator:

$$H\mathbf{x}(u) = \int_{\mathcal{T}} H(t,u)x(t)\,dt, \quad u \in \mathcal{U}.$$

The adjoint operator H^* is defined by

$$H^*\mathbf{y}(t) = \int_{\mathcal{U}} H(u,t)y(u)\,du, \quad t \in \mathcal{T}.$$

Therefore

$$\frac{\partial}{\partial x(t)}\|\mathbf{y} - H\mathbf{x}\|^2 = 2H^*(H\mathbf{x} - \mathbf{y})(t), \quad \forall t \in \mathcal{T}.$$

Returning to our problem of minimizing $\mathcal{E}(\mathbf{x})$, the stationarity condition takes the form of an Euler-Lagrange equation:

$$\begin{aligned} 0 &= \frac{\partial}{\partial x(t)}\mathcal{E}(\mathbf{x}) \\ &= \frac{\partial}{\partial x(t)}\|\mathbf{y} - H\mathbf{x}\|^2 + \lambda\,\frac{\partial}{\partial x(t)}\phi(\mathbf{x}) \\ &= 2H^*(H\mathbf{x} - \mathbf{y})(t) - \lambda\operatorname{div}\left(c(\|\nabla_t\, x(t)\|)\,\nabla_t\, x(t)\right), \quad t \in \mathcal{T}. \end{aligned} \tag{66}$$

The regularized solution $\hat{x}_{REG}(t)$ is the limit as $\tau \to \infty$ of the *time-varying image* $x(t,\tau)$ which satisfies the Partial Differential Equation (PDE) [3]

$$\begin{aligned} \frac{\partial x(t,\tau)}{\partial \tau} &= -\frac{\partial}{\partial x(t)}\mathcal{E}(x(\cdot,\tau)) \\ &= 2H^*(y - Hx)(t) - \lambda\,\frac{\partial}{\partial x(t,\tau)}\phi(x(\cdot,\tau)) \\ &= 2H^*(y - Hx)(t) + \lambda\operatorname{div}\left(c(\|\nabla_t\, x(t,\tau)\|)\,\nabla_t\, x(t,\tau)\right), \quad t \in \mathcal{T},\ \tau \in \mathbb{R}^+, \end{aligned} \tag{67}$$

$$\left.\frac{\partial x(t,\tau)}{\partial n}\right|_{\partial\mathcal{T}} = 0 \tag{68}$$

[3] This PDE is analogous to (62) if we view the iteration numbers k as discrete time instants and ϵ as a temporal sampling interval.

starting from some initial image $x(t,0)$. In (68), $\partial\mathcal{T}$ denotes the boundary of the image domain $\mathcal{T}$ and n denotes a vector normal to the boundary. The first term in the right side of (67) tries to force the solution to remain close to the data. The second term is a smoothing term which tends to minimize $\phi(\mathbf{x})$.

As discussed in the next subsection, (67) is also known as a diffusion equation. In order to reduce diffusion (smoothing) in the vicinity of edges, the diffusion coefficient $c(u)$ should decrease for large values of u. Diffusion has found applications to image segmentation and denoising. See [41, 42, 43], and the March 1998 special issue of the IEEE Transactions on Image Processing on PDEs and Geometry-Driven Diffusion in Image Processing and Analysis.

16.3 Connection to Diffusion Schemes

The diffusion scheme (67) may be viewed as an extension of the *anisotropic diffusion* scheme introduced by Perona and Malik [41], in which H was the identity operator. Their diffusion scheme is as follows:

$$\frac{\partial x(t,\tau)}{\partial \tau} = \operatorname{div}\left(c(\|\nabla_t\, x(t,\tau)\|)\,\nabla_t\, x(t,\tau)\right), \quad t \in \mathcal{T},\ \tau \in \mathbb{R}^+, \tag{69}$$

where the initial image $x(t,0)$ is the noisy image $y(t)$.

Comparing with (67), note that the data-fidelity term is omitted, and the PDE evolves in a way that progressively reduces the roughness penalty $\phi(\mathbf{x})$. While this scheme has no optimality property, it is a reasonable heuristic, is relatively simple to implement, and has led to some useful results in practice. The diffusion is terminated when the restored image "looks good enough".

Recalling (65), the special case where the diffusion coefficient $c = 1$ yields the classical *heat equation*

$$\frac{\partial x(t,\tau)}{\partial \tau} = \Delta^2 x(t,\tau). \tag{70}$$

The diffused image $x(t,\tau)$ is also the resulty of spatially convolving $x(t,0)$ with a 2-D isotropic Gaussian kernel of variance τ, as pointed out by Koenderink [44]. The heat equation is an isotropic diffusion scheme.

17 Bayesian Estimation

Unlike the maximum-likelihood and MAP estimation methods, which find a theoretical justification in an asymptotic setup, Bayesian estimation methods yield estimates that possess optimality properties for any sample size. The key ingredients of the Bayesian estimation technique are [4]:

- a prior $p(\mathbf{x})$, $\mathbf{x} \in \mathcal{X}$ and an observational model $p(\mathbf{y}|\mathbf{x})$, $\mathbf{y} \in \mathcal{Y}$;
- a family of estimators $\hat{\mathbf{x}}(\mathbf{y})$;
- a cost function $d(\mathbf{x}, \hat{\mathbf{x}})$ which quantifies the quality of any candidate estimator.

The Bayesian estimator is the one that minimizes the average cost over the family of candidate estimators.

Note the terminology: an *estimator* is an $\mathcal{X}$-valued function defined over the domain $\mathcal{Y}$; an *estimate* is a specific value of $\mathbf{x} \in \mathcal{X}$ returned by the function for a fixed input $\mathbf{y}$.

Some possible choices for the cost function are:

- Mean Squared Error (MSE): $d(\mathbf{x}, \hat{\mathbf{x}}) = \|\mathbf{x} - \hat{\mathbf{x}}\|^2$
- Frequency-weighted MSE: $d(\mathbf{x}, \hat{\mathbf{x}}) = \|\mathbf{h} \star (\mathbf{x} - \hat{\mathbf{x}})\|^2$ where $\mathbf{h}$ is a linear shift-invariant filter, and $\star$ denotes convolution. The frequency reponse of the filter is designed to approximate the frequency response of the Human Visual System.
- Perceptual error measures: these are nonquadratic distortion measures, taking local image activity, texture, masking effects, etc. into account.

We now define the following quantities:

- Risk:
$$r(\hat{\mathbf{x}}) \triangleq \mathbb{E}_{\mathbf{XY}}[d(\mathbf{X}, \hat{\mathbf{x}}(\mathbf{Y}))]$$
is the cost averaged over all possible values of $\mathbf{X}$ and $\mathbf{Y}$.
- Conditional risk:
$$R(\hat{\mathbf{x}}|\mathbf{x}) \triangleq \mathbb{E}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}[d(\mathbf{x}, \hat{\mathbf{x}}(\mathbf{Y}))]$$
is the cost averaged over $\mathbf{Y}$ and conditioned on $\mathbf{X} = \mathbf{x}$.
- Posterior risk:
$$R(\hat{\mathbf{x}}|\mathbf{y}) \triangleq \mathbb{E}_{\mathbf{X}|\mathbf{Y}=\mathbf{y}}[d(\mathbf{X}, \hat{\mathbf{x}}(\mathbf{y}))]$$
is the cost averaged over $\mathbf{X}$ and conditioned on $\mathbf{Y} = \mathbf{y}$.

The conditional risk is useful to assess the average (over all possible data $\mathbf{y}$) performance of an estimator on a given image $\mathbf{x}$. The posterior risk is useful to assess the average (over all possible images $\mathbf{x}$) performance of an estimator on a given dataset $\mathbf{y}$. By averaging the conditional (resp. posterior) risk over $\mathbf{X}$ (resp. $\mathbf{Y}$), we obtain the Bayes risk:

$$r(\hat{\mathbf{x}}) = \mathbb{E}_{\mathbf{X}}[R(\hat{\mathbf{x}}|\mathbf{X})] = \mathbb{E}_{\mathbf{Y}}[R(\hat{\mathbf{x}}|\mathbf{Y})].$$

The Bayes estimator $\hat{\mathbf{x}}_B$ is the one that minimizes the risk $r(\hat{\mathbf{x}})$ over all possible estimators $\hat{\mathbf{x}}$. For any given $\mathbf{y}$, the estimate $\hat{\mathbf{x}}(\mathbf{y})$ is obtained by minimizing the posterior risk $R(\hat{\mathbf{x}}|\mathbf{y})$ over all possible $\hat{\mathbf{x}}$. Indeed,

$$r(\hat{\mathbf{x}}) = \int R(\hat{\mathbf{x}}|\mathbf{y})\, p(\mathbf{y})\, d\mathbf{y}$$

and therefore

$$\min_{\hat{\mathbf{x}}(\cdot)} r(\hat{\mathbf{x}}) = \int \left[\min_{\hat{\mathbf{x}}(\mathbf{y})} R(\hat{\mathbf{x}}|\mathbf{y})\right] p(\mathbf{y})\, d\mathbf{y}$$

i.e., the optimization problems involving different $\mathbf{y}$ are decoupled. The estimate $\hat{\mathbf{x}}(\mathbf{y})$ is the minimizer of the integral

$$R(\hat{\mathbf{x}}|\mathbf{y}) = \int d(\mathbf{x}, \hat{\mathbf{x}})\, p(\mathbf{x}|\mathbf{y})\, d\mathbf{x}.$$

While difficult to evaluate in general (due to the integration over a high-dimensional space), this expression shows the fundamental role played by the posterior distribution $p(\mathbf{x}|\mathbf{y})$. The posterior distribution plays a fundamental role in all Bayesian inference problems.

Let us now derive the Bayesian estimator for the MSE cost function. The resulting estimator is the so-called Minimum Mean Squared Error (MMSE) estimator and will be denoted by $\hat{\mathbf{x}}_{MMSE}$. We need to minimize the posterior risk

$$R(\hat{\mathbf{x}}|\mathbf{y}) = \int \|\mathbf{x} - \hat{\mathbf{x}}\|^2\, p(\mathbf{x}|\mathbf{y})\, d\mathbf{x}$$

over $\hat{\mathbf{x}} \in \mathcal{X}$. This is a quadratic optimization problem which can be solved by setting the gradient of the posterior risk over $\hat{\mathbf{x}}$. This yields

$$\begin{aligned} 0 &= \nabla_{\hat{\mathbf{x}}} \int \|\mathbf{x} - \hat{\mathbf{x}}\|^2\, p(\mathbf{x}|\mathbf{y})\, d\mathbf{x} \\ &= 2 \int (\hat{\mathbf{x}} - \mathbf{x})\, p(\mathbf{x}|\mathbf{y})\, d\mathbf{x} \\ &= \hat{\mathbf{x}} - \int \mathbf{x}\, p(\mathbf{x}|\mathbf{y})\, d\mathbf{x}. \end{aligned}$$

Hence the MMSE estimator is the conditional mean:

$$\hat{\mathbf{x}}_{MMSE}(\mathbf{y}) = \int \mathbf{x}\, p(\mathbf{x}|\mathbf{y})\, d\mathbf{x}.$$

Next, we evaluate the Bayesian estimator for the so-called zero/one cost function:

$$d(\mathbf{x}, \hat{\mathbf{x}}) = 1_{\{\hat{\mathbf{x}} \notin \mathcal{B}(\mathbf{x};\epsilon)\}}$$

where

$$\mathcal{B}(\mathbf{x};\epsilon) \triangleq \{\mathbf{x}' \ : \ \|\mathbf{x}' - \mathbf{x}\| \leq \epsilon\}$$

is the sphere of radius ϵ centered at $\mathbf{x}$. The zero/one cost function assigns zero cost if the estimate is within distance ϵ of the true $\mathbf{x}$, and assigns a fixed cost of 1 otherwise (irrespective of the magnitude of the error).

The posterior risk takes the form

$$\begin{aligned} R(\hat{\mathbf{x}}|\mathbf{y}) &= \int 1_{\{\hat{\mathbf{x}} \notin \mathcal{B}(\mathbf{x};\epsilon)\}} \, p(\mathbf{x}|\mathbf{y}) \, d\mathbf{x} \\ &= \int 1_{\{\mathbf{x} \notin \mathcal{B}(\hat{\mathbf{x}};\epsilon)\}} \, p(\mathbf{x}|\mathbf{y}) \, d\mathbf{x} \\ &= 1 - \int_{\mathcal{B}(\hat{\mathbf{x}};\epsilon)} p(\mathbf{x}|\mathbf{y}) \, d\mathbf{x}. \end{aligned}$$

In the limit as $\epsilon \to 0$, the integral tends to the volume of the sphere times $p(\hat{\mathbf{x}}|\mathbf{y})$. Therefore, minimizing the posterior risk is equivalent to finding the mode of the posterior distribution. This is the same as the MAP estimation criterion:

$$\hat{\mathbf{x}}_{MAP}(\mathbf{y}) = \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}).$$

To illustrate the dependency of Bayesian estimators with respect to the cost function, consider the following example.

Problem: Let x be the parameter of the exponential distribution $p(y|x) = x\, e^{-xy}\, 1_{\{y \geq 0\}}$. Assume an exponential prior, $p(x) = a\, e^{-ax}\, 1_{\{x \geq 0\}}$, where a is a given constant. Find the MAP and MMSE estimators of X given Y.

Solution:

Step 1: derive the posterior $p(x|y)$. The joint pdf is given by

$$p(x, y) = p(y|x)p(x) = ax\, e^{-(a+y)x}, \quad x, y \geq 0.$$

Integrating over x, we obtain

$$p(y) = \int_0^\infty ax\, e^{-(a+y)x}\, dx = \frac{a}{(a+y)^2}, \quad y \geq 0.$$

Hence the posterior is given by

$$p(x|y) = \frac{p(x, y)}{p(y)} = (a+y)^2 x\, e^{-(a+y)x}, \quad x \geq 0.$$

This is a unimodal distribution, taking value 0 at $x = 0$.

Step 2: the MAP estimator is obtained from the equation

$$0 = \frac{d}{dx} \ln p(x|y) = \frac{d}{dx}[\ln x - (a+y)x] = \frac{1}{x} - (a+y).$$

Hence

$$\hat{x}_{MAP}(y) = \frac{1}{a+y}.$$

Note that $\hat{x}_{MAP}(y)$ could have been calculated more straightforwardly, bypassing evaluation of the marginal $p(y)$.

Step 3: the MMSE estimator is given by

$$\hat{x}_{MMSE}(y) = \int_0^\infty x\, p(x|y)\, dx = \int_0^\infty (a+y)^2 x^2\, e^{-(a+y)x}\, dx = \frac{2}{a+y}.$$

The MAP and MMSE estimators have the same form (inversely proportional to $a + y$), but the proportionality constants involved are different. This simple example suggests that in image restoration also, the choice of the distortion measure $d(\mathbf{x}, \hat{\mathbf{x}})$ may have an impact on the Bayesian estimator. Limited empirical studies suggest that this is true to some extent. The MMSE estimator is generally somewhat better than the MAP estimator, but not dramatically better.

18 Applications (II)

In addition to the classical applications listed in Section 12, we will consider two other applications in some detail. Both have received significant attention in the last five years.

18.1 Image Inpainting

This problem originated in the context of restoration of old, damaged pictures, photographic film, paintings, and other artwork [45]. The image degradation takes the form of creases, streaks, wrinkles, and possibly small stains. The mathematical model is that the corrupted region $\mathcal{R}$ is a narrow area (or the union of narrow areas). The observed pixels agree with the original pixels over the whole image, except for the region $\mathcal{R}$. Inside $\mathcal{R}$, the observed pixels convey no information whatsoever about the original pixels.

The corresponding statistical model would be

$$y_i \begin{cases} = x_i & : \ i \notin \mathcal{R} \\ \text{independent of } \mathbf{x} & : \ i \in \mathcal{R}. \end{cases} \tag{71}$$

High-quality restoration of the original $\mathbf{x}$ should be possible if the region $\mathcal{R}$ is narrow and if information outside $\mathcal{R}$ can be successfully propagated into $\mathcal{R}$.

Assuming an image prior $p(\mathbf{x})$, the MAP estimation criterion takes the form

$$\hat{\mathbf{x}}_{MAP} = \arg \max_{\mathbf{x} \,:\, \mathbf{x}_{\mathcal{R}^c} = \mathbf{y}_{\mathcal{R}^c}} p(\mathbf{x})$$

where $\mathcal{R}^c$ denotes the complementary (clean) region. For MRF models of the form $p(\mathbf{x}) \propto \exp\{-\mathcal{E}(\mathbf{x})\}$, the problem above may be written as

$$\hat{\mathbf{x}}_{MAP} = \arg \min_{\mathbf{x} \,:\, \mathbf{x}_{\mathcal{R}^c} = \mathbf{y}_{\mathcal{R}^c}} \mathcal{E}(\mathbf{x})$$

and solved using anisotropic diffusion:

$$\hat{\mathbf{x}}_{\mathcal{R}}^{(k+1)} = \hat{\mathbf{x}}_{\mathcal{R}}^{(k)} - \epsilon \, \nabla_{\mathbf{x}_{\mathcal{R}}} \mathcal{E}(\mathbf{x})|_{\mathbf{x}=\hat{\mathbf{x}}^{(k)}} \,,$$

where k denotes iteration number.

18.2 Superresolution

Superresolution refers to the problem of constructing a high-resolution image given a sequence of low-resolution images. Applications include video surveillance, where the camera has poor resolution and contrast, and it is desired to enhance the picture of a person of interest that appears in the scene [46, 47].

Image superresolution may be viewed as an extension of the classical image interpolation problem where the resolution of a *single image* should be improved. There are may classical algorithms for image interpolation, notably those based on linear splines and cubic splines. These algorithms use simple linear filters and do not improve the resolution – they just produce an image with more pixels and a blurry appearance. To achieve true resolution enhancement one could use an advanced image prior, formulate an observational model, and use MAP estimation to obtain a solution.

19 Statistical Image Inference Based on Graphical Models

Consider the following simple example [4]: the original image $\mathbf{x}$ follows an Ising model,

$$p(\mathbf{x}) = \frac{1}{Z(\alpha, \beta)} \exp \left\{ -\alpha \sum_i x_i - \beta \sum_{i \sim j} x_i x_j \right\}$$

and the observational model is iid:

$$p(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{N} p(y_i|x_i) = \exp \left\{ \sum_{i=1}^{N} \ln p(y_i|x_i) \right\}.$$

[4] More realistic models will be considered below.

Find

$$\hat{\mathbf{x}}_{MAP} = \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y})$$

or

$$\hat{\mathbf{x}}_{MMSE} = \sum_{\mathbf{x}} \mathbf{x}\, p(\mathbf{x}|\mathbf{y})$$

where the posterior distribution takes the form

$$p(\mathbf{x}|\mathbf{y}) = \frac{1}{Z(\alpha,\beta)\, p(\mathbf{y})} \exp\left\{ \left[-\alpha \sum_i x_i + \ln p(y_i|x_i) \right] - \beta \sum_{i\sim j} x_i x_j \right\}.$$

The corresponding graphical model is shown in Fig. 16.

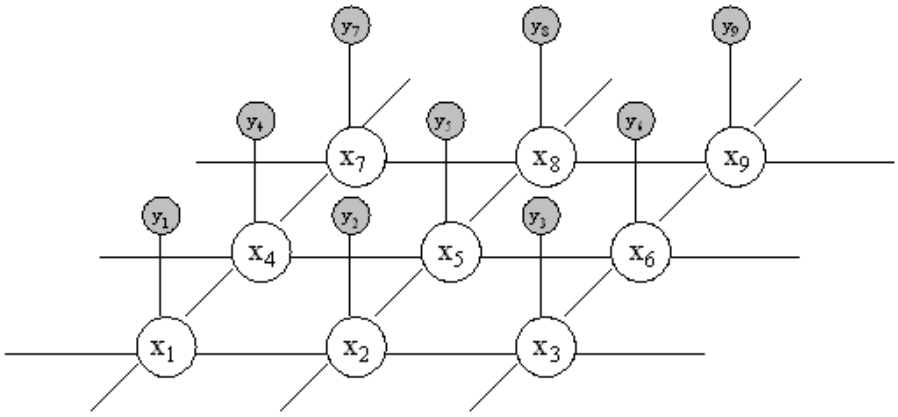


Figure 16: A graphical model for image restoration: $\{y_i\}$ are the data, $\{x_i\}$ are the hidden variables to be estimated.

19.1 Exponential Families

In the above Ising example, $p(\mathbf{x}|\mathbf{y})$ takes an exponential form. More generally, this is always the case when the *joint ditribution* for $(\mathbf{x}, \mathbf{y})$ takes an exponential form:

$$p(\mathbf{x}, \mathbf{y}|\theta) = q(\mathbf{x}, \mathbf{y}) \exp\{\boldsymbol{\theta}^T \mathbf{u}(\mathbf{x}, \mathbf{y}) - A(\boldsymbol{\theta})\}.$$

Indeed, the posterior distribution takes the form

$$p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}) = \frac{p(\mathbf{x}, \mathbf{y}|\theta)}{\int p(\mathbf{x}, \mathbf{y}|\theta)\, d\mathbf{x}} = \frac{q(\mathbf{x}, \mathbf{y}) \exp\{\boldsymbol{\theta}^T \mathbf{u}(\mathbf{x}, \mathbf{y})\}}{\int q(\mathbf{x}, \mathbf{y}) \exp\{\boldsymbol{\theta}^T \mathbf{u}(\mathbf{x}, \mathbf{y})\}\, d\mathbf{x}}.$$

Writing

$$A_{\mathbf{y}}(\boldsymbol{\theta}) \triangleq \ln \int q(\mathbf{x},\mathbf{y}) \exp\{\boldsymbol{\theta}^T \mathbf{u}(\mathbf{x},\mathbf{y})\}\, d\mathbf{x}$$

we see that the posterior pdf takes the form

$$p(\mathbf{x}|\mathbf{y},\boldsymbol{\theta}) = q(\mathbf{x},\mathbf{y}) \exp\{\boldsymbol{\theta}^T \mathbf{u}(\mathbf{x},\mathbf{y}) - A_{\mathbf{y}}(\boldsymbol{\theta})\}$$

i.e., is also in an exponential family.

The solution to the MAP estimation problem above involves minimization of a (possibly nonconvex) energy function $\mathcal{E}_{\mathbf{y}}(\mathbf{x})$. The solution to the MMSE estimation problem seems even harder due to the integration over all possible configurations $\mathbf{x}$. The solution proposed by Geman and Geman [8] was simulated annealing, which is computationally costly. Below, we consider belief propagation as a computationally attractive alternative.

19.2 Exact Inference

Observe that the MAP and MMSE estimators can be written as

$$\begin{aligned} \hat{x}_{MAP,i} &= \max_{x_i} q(x_i|\mathbf{y}) \\ \hat{x}_{MMSE,i} &= \sum_{x_i} x_i\, p(x_i|\mathbf{y}), \quad i \in \mathcal{T} \end{aligned}$$

where

$$p(x_i|\mathbf{y}) = \sum_{\mathbf{x}' \,:\, x'_i = x_i} p(\mathbf{x}|\mathbf{y})$$

is the marginal posterior distribution for x_i given $\mathbf{y}$, and

$$q(x_i|\mathbf{y}) = \max_{\mathbf{x}' \,:\, x'_i = x_i} p(\mathbf{x}|\mathbf{y})$$

is generally not a pdf (does not sum to 1).

For a tree-structured graph, the MAP and MMSE estimators can be computed exactly using the max-product and sum-product algorithms of Chapter III.7. Assume the prior for $\mathbf{x}$ takes the factorial form

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{i \sim j} \psi_{ij}(x_i, x_j).$$

The observational model takes the product form

$$p(\mathbf{y}|\mathbf{x}) = \prod_{i \in \mathcal{T}} \phi(x_i, y_i).$$

Therefore the joint distribution is of the form

$$p(\mathbf{x},\mathbf{y}) = \frac{1}{Z} \prod_i \phi(x_i, y_i) \prod_{j \sim i} \psi_{ij}(x_i, x_j).$$

For message-passing algorithms, the presence of $\mathbf{y}$ does not introduce any special difficulty. Consider the MAP estimation problem for instance. By choosing an appropriate elimination order, we can generate the message from any node i to an adjacent node j as follows:

$$m_{ij}(x_j) = \max_{x_i} \phi(x_i, y_i)\psi_{ij}(x_i, x_j) \prod_{k \in \mathcal{N}(i)\setminus\{j\}} m_{ki}(x_i) \tag{72}$$

$$q(x_i|\mathbf{y}) \propto \phi(x_i, y_i) \prod_{k \in \mathcal{N}(i)} m_{ki}(x_i). \tag{73}$$

(where initial values of the messages are chosen equal to 1.) Interestingly, the messages can also be computed by implementing the following recursion:

$$m_{ij}^{\tau}(x_j) = \max_{x_i} \phi(x_i, y_i)\psi_{ij}(x_i, x_j) \prod_{k \in \mathcal{N}(i)\setminus\{j\}} m_{ki}^{\tau-1}(x_i) \tag{74}$$

where $\tau = 1, 2, \cdots$ denotes iteration number. The recursion is guaranteed to converge after a finite number of iterations [48]. Once the messages are available at all nodes, the MAP estimates can be simply obtained.

For Hidden Markov Models (1D), this algorithm coincides with the forward-backward estimation algorithm [49].

For MMSE estimation we have

$$m_{ij}(x_j) = \sum_{x_i} \phi(x_i, y_i)\psi_{ij}(x_i, x_j) \prod_{k \in \mathcal{N}(i)\setminus\{j\}} m_{ki}(x_i) \tag{75}$$

$$p(x_i|\mathbf{y}) = \frac{1}{Z}\phi(x_i, y_i) \prod_{k \in \mathcal{N}(i)} m_{ki}(x_i). \tag{76}$$

19.3 Approximate Inference

For graphs that contain cycles, an elimination procedure could in principle be used to perform exact inference. Unfortunately the complexity of that procedure grows exponentially with the size of the graph. A simple but surprisingly effective approach is to run the iterative algorithm (74) and use the resulting messages to compute the MAP or MMSE estimates as was done in the previous section. Unlike acyclic graphs, the recursion is not guaranteed to converge, and even if convergence occurs, the resulting estimates are not guaranteed to coincide with the exact MAP or MMSE estimates.

This approach is also known as *loopy belief propagation* and may be expected to work well if the graph is "nearly loop-free". Even when the graph contains many loops (e.g., the Ising model), loopy belief propagation often gives excellent results in practice, as will be evidenced in the next sections. At a theoretical level, loopy belief propagation enjoys an elegant variational interpretation [50, 51, 52].

An even simpler method for approximate inference is *mean-field theory*, which was popular in the MRF image restoration comunity during the 1990's, and also admits a variational interpretation [50, 51, 52].

20 Parametric Prior Estimation

As discussed in Section 13, exponential families provide a flexible and computationally attractive framework for modeling images. Assume a model of the form

$$p(\mathbf{x}) = \exp\{\boldsymbol{\theta}^T \mathbf{u}(\mathbf{x}) - A(\boldsymbol{\theta})\}$$

where $\boldsymbol{\theta} \in \Theta$ is not fixed (unlike in the previous section) but must be estimated from *training data* $\mathbf{x}^1, \cdots, \mathbf{x}^K$, assumed to be representative of the actual image to be reconstructed.

Recall (46): the likelihood equation for the *mean-value parameter* $\boldsymbol{\xi} = \nabla A(\boldsymbol{\theta})$ yields simply

$$\hat{\boldsymbol{\xi}}_{ML} = \frac{1}{K} \sum_{i=1}^{K} \mathbf{u}(\mathbf{x}^i).$$

Then $\hat{\boldsymbol{\theta}}_{ML}$ may be obtained by inverting the nonlinear transformation $\nabla A(\cdot)$.

A closely related approach has ben adopted by Roth and Black [53] to fit homogeneous MRF models with clique system $\mathscr{C}$ and nonconvex potential functions of the form

$$V_{\mathcal{C}}(\mathbf{x}_{\mathcal{C}}) = \sum_{i \in \mathcal{I}} \varphi(\mathbf{J}_i \cdot \mathbf{x}_{\mathcal{C}}, \alpha_i)$$

where $\varphi(u, \alpha) = \alpha \log\left(1 + \frac{u^2}{2}\right)$, $\alpha \geq 0$, and $\mathbf{J}_i$, $i \in \mathcal{I}$, are basis vectors with the same dimension as $\mathcal{C}$. Taking the dot product of $\mathbf{x}_{\mathcal{C}}$ with $\mathbf{J}_i$ may be thought of as a local filtering operation. The number of filters, $|\mathcal{I}|$, could be lower, equal, or greater than the size $|\mathcal{C}|$ of the cliques.

In all cases, the image prior can be written as

$$p(\mathbf{x}) = \exp\left\{ -\ln Z(\mathsf{J}, \boldsymbol{\alpha}) - \sum_{\mathcal{C} \in \mathscr{C}} \sum_{i \in \mathcal{I}} \varphi(\mathbf{J}_i \cdot \mathbf{x}_{\mathcal{C}}, \alpha_i) \right\}. \tag{77}$$

In their model (77), which they call *field of experts*, the cliques have size 5×5, there are $|\mathcal{I}| = 24$ filters, and the parameters $\boldsymbol{\alpha} = \{\alpha_i \, i \in \mathcal{I}\}$ and $\mathsf{J} = \{J_i, \, i \in \mathcal{I}\}$ are estimated from training data $\mathbf{x}^1, \cdots, \mathbf{x}^K$. The ML parameter estimation problem takes the form

$$\min_{\mathsf{J}, \boldsymbol{\alpha}} \left\{ \mathcal{E}(\mathsf{J}, \boldsymbol{\alpha}) \triangleq \ln Z(\mathsf{J}, \boldsymbol{\alpha}) + \sum_{\mathcal{C} \in \mathscr{C}} \sum_{i \in \mathcal{I}} \sum_{k=1}^{K} \varphi(\mathbf{J}_i \cdot \mathbf{x}_{\mathcal{C}}^k, \alpha_i) \right\}$$

and is solved using an approximate gradient descent algorithm. The number of unknowns is equal to $(24 \times 25) + 24 = 624$, so the optimization problem above is challenging. The training image set is represented as a set of 60,000 image patches.

21 Nonparametric Prior Estimation

A completely different approach to the problem of estimating image priors is based on concepts from the statistical literature on probability density estimation. One does not assume any parametric form for $p(\mathbf{x})$ but estimates the whole distribution from training data. As discussed in Chapter II.2, the huge dimensionality of $\mathbf{x}$ poses considerable challenges. Let us examine some basic ideas.

21.1 Discrete Priors

Ignoring momentarily the dimensionality problem, assume we are given K examples $\mathbf{x}^1, \cdots, \mathbf{x}^K$ drawn from $p(\mathbf{x})$ and estimate $p(\mathbf{x})$ as a normalized sum of Dirac impulses located at each $\mathbf{x}^k$:

$$p(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^{K} \delta(\mathbf{x} - \mathbf{x}^k). \tag{78}$$

Let $\Omega = \{\mathbf{x}^1, \cdots, \mathbf{x}^K\}$ denote the training set.

While this discrete prior seems to be an unrealistic model (precluding the possibility that any future $\mathbf{x}$ could be anything else than an element of Ω), this may be good enough for some inference problems. Consider the denoising problem with additive white Gaussian noise. The MAP estimator under the discrete prior model takes the form

$$\hat{\mathbf{x}}_{MAP} = \arg \min_{\mathbf{x}^k \in \Omega} \|\mathbf{y} - \mathbf{x}^k\|$$

which is simply called the *nearest-neighbor estimator* and could conceivably be an adequate estimator if the training set was vast enough.

An important refinement on this idea is to assign nonuniform weights to the examples $\mathbf{x}^k$. The prior would be of the form

$$p(\mathbf{x}) = \sum_{k=1}^{K} w_k \, \delta(\mathbf{x} - \mathbf{x}^k) \tag{79}$$

where $\{w_k\}$ are nonnegative weights summing to 1. If each w_k is a negative power of 2, of the form $2^{-L(k)}$, one may think of $\tilde{\Omega} = \{\mathbf{x}^k, L(k), 1 \leq k \leq K\}$ as a *codebook* whose elements $\mathbf{x}^k$ are represented using variable-length codes. Then (79) is an instance of a *complexity prior*, and fundamental bounds on estimation performance can be derived for such priors [54]. The MAP (or complexity-regularized) estimator for the denoising problem takes the form

$$\hat{\mathbf{x}}_{MAP} = \arg \min_{\mathbf{x}^k \in \Omega} \left[\frac{\log_2 e}{2\sigma^2} \|\mathbf{y} - \mathbf{x}^k\|^2 + L(k) \right].$$

This is consistent with the intuitive notion that unlikely images should be assigned long codewords in a codebook. Of course it remains to design $\tilde{\Omega}$ in an appropriate way. One approach is to use codebooks derived from JPEG and wavelet compression algorithms [55].

Another approach, which is more recent and rather interesting, is based on training for *image patches* [56]. Partition the image $\mathbf{x}$ into a set of square patches $\{\mathbf{x}_i\}$ (of size, say, 7×7). From the image training set extract patches to form a *patch training set* Ω_p consisting of tens of thousands of image patches. The prior model is constrained to be of the form

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{i \sim j} \psi_{ij}(\mathbf{x}_i, \mathbf{x}_j) \tag{80}$$

where the compatibility functions are now estimated nonparametrically from the training data. Specifically,

$$\psi_{ij}(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} \exp\left\{-\frac{d(\mathbf{x}_i, \mathbf{x}_j)}{2\sigma_0^2}\right\} & : \ \mathbf{x}_i, \mathbf{x}_j \in \Omega_p \\ 0 & : \ \text{else} \end{cases}$$

where $d(\mathbf{x}_i, \mathbf{x}_j)$ is a Euclidean measure of similarity between patches $\mathbf{x}_i$ and $\mathbf{x}_j$.

This prior model assigns positive probability only to images $\mathbf{x}$ whose patches agree with a training patch from Ω_p. Hence $p(\mathbf{x})$ is a discrete prior of the form (79), where the probabilities w_k are functions of the similarities between adjacent patches, and Ω is the N_p-fold Cartesian product of Ω_p, where N_p is the number of patches in the image. If the resulting $\{w_k\}$ were negative powers of 2, we could formally view $p(\mathbf{x})$ as a complexity prior.

The observational model takes the product form

$$p(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{N_p} \phi(\mathbf{x}_i, \mathbf{y}_i).$$

The paper [56] uses several variations on the prior model (80) and estimates $\mathbf{x}$ given $\mathbf{y}$ using loopy belief propagation. Excellent results have been reported for image denoising and superresolution applications. The performance of the statistical inference algorithms is fairly insensitive to the choice of the training set. Choosing quirky training sets (made of iid noise or other unnatural patterns) would however negatively impact the performance of the algorithms, because of the reduced likelihood to find training patches that are a good match to the data.

21.2 Kernel-Based Estimation

If one takes the view that training data merely indicate the regions of probability space that have significant probability, a possible improvement over the discrete prior (79) is the kernel estimator

$$p(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^{K} q(\mathbf{x} - \mathbf{x}^k) \tag{81}$$

where the *kernel* $q(\mathbf{x})$ is a pdf. One may think of (81) as the convolution of the discrete prior (79) with the kernel, i.e., a smoothing of the discrete prior.

Baker and Kanade [46] extend this idea in an image modeling context. Roughly speaking, they view $\mathbf{x}$ as made of two components: a classification index c, and a detail component $\mathbf{d}$

modeled as Gaussian with mean $\mathbf{x}_c$ and covariance matrix R_c. Equivalently, their model is $\mathbf{x} = \mathbf{x}_c + \mathbf{d}$, or

$$p(\mathbf{x}) = \sum_c p(c)\, q_c(\mathbf{x} - \mathbf{x}_c)$$

where $q_c = \mathcal{N}(0, R_c)$.

The classification index is estimated from the data $\mathbf{y}$ using a classification algorithm which is outlined below. Baker and Kanade study a superresolution problem in which the data $\mathbf{y}$ are related to $\mathbf{x}$ via a linear Gaussian observational model. Conditioned on c, the problem of recovering $\mathbf{d}$ from $\mathbf{y}$ is a linear least-squares problem.

The classification index c, the mean image $\mathbf{x}_c$, and the covariance matrix R_c are obtained as follows. A training set $\Omega = \{\mathbf{x}^1, \cdots, \mathbf{x}^K\}$ is given, and for each image $\mathbf{x}^k \in \Omega$, a Gaussian image pyramid, a Laplacian pyramid, and three derivative pyramids are constructed. To any pixel at any given resolution level in the image pyramid, a *parent structure vector* is assigned, consisting of the values of the pixel in the five pyramids and their parents. Similarly, a parent structure vector is associated with the pixels of the image data $\mathbf{y}$. The classification algorithm then seeks the *best match* between each observed parent structure vector and a corresponding parent structure vector from the training set. (A different match is generally obtained for each pixel.) The image formed by aggregating the best matching pixels is called $\mathbf{x}_c$. The covariance matrix R_c, which is intended to capture second-order statistics of the detail image $\mathbf{d}$, is defined in terms of the local gradients of $\mathbf{x}_c$. As can be seen from the description of the algorithm, the number of possible values for the classification index c is exponential in the size of the image $\mathbf{y}$ – much larger than the size of the training set.

The name coined by Baker and Kanade for their superresolution algorithm is the *hallucination algorithm*. Excellent results have been reported in [46] for fairly generic training sets. The algorithm has found promising applications to face recognition, and a natural training set in such applications would be made of human faces. The "hallucination" label is sometimes fitting in the sense that the algorithm tends to output images that are closely related to images in the training set. In particular, exceptional results are obtained when the test face image $\mathbf{y}$ corresponds to an individual already in the training set. However, if $\mathbf{y}$ is unrelated to a human face (e.g., $\mathbf{y}$ is a constant background), the algorithm tends to "hallucinate" the outline of a face.

References

[1] A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice-Hall, 1989.

[2] D. L. Snyder and M. I. Miller, *Random Point Processes in Time and Space*, 2nd Ed., Springer-Verlag, New York, 1991.

[3] K. Lange and J. Fessler, "Globally Convergent Algorithms for Maximum A Posteriori Transmission Tomography," *IEEE Trans. Im. Proc.*, Vol. 4, No. 10, pp. 1430—1438, Oct. 1995.

[4] H. V. Poor, *An Introduction to Signal Detection and Estimation*, Springer-Verlag, 1994.

[5] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, (with discussion) J. Roy. Stat. Soc. B., Vol. , No. 1, pp. 1-38, 1977.

[6] C. F. J. Wu, "On the Convergence Properties of the EM Algorithm," *Ann. Stat.*, Vol. 11, No. 1, pp. 95—103, 1983.

[7] J. A. Fessler and A. O. Hero, "Space-Alternating Generalized Expectation-Maximization Algorithm," *IEEE Trans. on Sig. Proc.*, Vol. 42, No. 10, pp. 2664—2677, Oct. 1994.

[8] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. on PAMI*, Vol. 6, No. 6, pp. 721—741, Nov. 1984.

[9] J. Besag, "Towards Bayesian Image Analysis," *J. Applied Statistics*, Vol. 16, No. 3, pp. 395-407, 1989.

[10] M. R. Banham and A. K. Katsaggelos, "Digital Image Restoration," *IEEE Sig. Proc. Magazine*, pp. 24—41, March 1997.

[11] J. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd Ed., Springer Verlag, 1985.

[12] D. Luenberger, *Optimization by Vector Space Methods*, Wiley, 1969.

[13] L. I. Rudin, S. Osher and E. Fatemi, "Nonlinear Total Variation Based Noise Removal Algorithms," *Physica D*, Vol. 60, pp. 259—268, 1992.

[14] S. Geman and G. Reynolds, "Constrained Restoration and Recovery of Discontinuities," *IEEE Trans. on PAMI*, Vol. 14, No. 3, pp. 367—383, March 1992.

[15] C. Bouman and K. Sauer, "A Generalized Gaussian Image Model for Edge-Preserving MAP Estimation," *IEEE Trans. Im. Proc.*, Vol. 2, No. 3, pp. 296—310, July 1993.

[16] P. Charbonnier, L. Blanc-Féraud, G. Aubert and M. Barlaud,, "Deterministic Edge-Preserving Regularization in Computed Imaging," *IEEE Trans. Im. Proc.*, Vol. 6, No. 2, pp. 298—311, Feb. 1997.

[17] M. Nikolova, "Regularisation Functions and Estimators," *Proc. ICIP96*, Vol. 2, pp. 457—460, Lausanne, Switzerland, 1996.

[18] M. Belge, M. Kilmer and E. Miller, "Wavelet Domain Image Restoration with Adaptive Edge-Preserving Regularization." *IEEE Trans. Image Processing*, Vol. 9, No. 4, pp. 597—608, Aug. 2000.

[19] J. Liu and P. Moulin, "Complexity-Regularized Image Restoration," *Proc. IEEE Int. Conf. on Image Proc. (ICIP98)*, Vol. 1, pp. 555—559, Chicago, Oct. 1998.

[20] P. Moulin and J. Liu, "Analysis of Multiresolution Image Denoising Schemes Using Generalized-Gaussian and Complexity Priors," *IEEE Trans. on Information Theory*, Special Issue on Multiscale Analysis, Vol. 45, No. 3, pp. 909—919, Apr. 1999.

[21] P. L. Combettes, "The Foundations of Set-Theoretic Estimation," *Proc. IEEE*, Vol.81, No. 2, pp. 182—207, 1993.

[22] D. M. Titterington, "General Structure of Regularization Procedures in Image Reconstruction," *Astron. Astrophysics*, 144, pp. 381-387, 1985.

[23] A. N. Thompson, J. C. Brown, J. W. Kay and D. N. Titterington, "A Study of Methods of Choosing the Smoothing Parameter in Image Restoration by Regularization," *IEEE Trans. on PAMI*, Vol. 13, No. 4, pp. 326—, April 1991.

[24] G. Archer and D. M. Titterington, "On Some Bayesian/Regularization Methods for Image Restoration," *IEEE Trans. Im. Proc.*, Vol. 4, No. 7, pp. 989-995, July 1995.

[25] T. P. O'Rourke and R. L. Stevenson, "Improved Image Decompression for Reduced Transform Coding Artifacts," *IEEE Trans. CSVT*, Vol. 5, No. 6, pp. 490-499, Dec. 1995.

[26] M. A. Robertson and R. L. Stevenson, "Reducing the Computational Complexity of a MAP Post-Processing Algorithm for Video Sequences," *Proc. IEEE Int. Conf. on Image Proc. (ICIP98)*, Vol. 1, Chicago, Oct. 1998.

[27] P. Salama, N. B. Shroff and E. J. Delp, "Error Concealment in Encoded Video Streams," in *Signal Recovery Techniques for Image and Video Compression and Transmission*, Eds. N. Galatsanos and A. K. Katsaggelos, Kluwer, Boston, 1998.

[28] R. L. Stevenson, "Inverse Halftoning via MAP Estimation," *IEEE Trans. Im. Proc.*, Vol. 6, No. 4, pp. 574-583, Apr. 1997.

[29] K. L. Lagendijk and J. Biemond, *Iterative Identification and Restoration of Images*, Kluwer, Boston, 1991.

[30] Y.-L. You and M. Kaveh, "A Regularization Approach to Joint Blur Identification and Image Restoration," *IEEE Trans. Im. Proc.*, Vol. 5, No. 3, pp. 416-428, March 1996.

[31] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, Springer, New York, 1998.

[32] E. T. Jaynes, "Information Theory and Statistical Mechanics," *Physics Review*, Vol. 106, pp. 620—630, 1957.

[33] E. T. Jaynes, "On the Rationale of Maximum-Entropy Methods," *Proceedings of the IEEE*, Vol. 70, No. 9, pp. 939—952, Sep. 1982.

[34] S. C. Zhu, Y. N. Wu and D. Mumford, "Minimax Entropy Principle and Its Application to Texture Modeling," *Neural Computation*, Vol. 9, pp. 1627—1660, 1997.

[35] S. C. Zhu and D. Mumford, "Prior Learning and Gibbs Reaction-Diffusion," *IEEE Trans. PAMI*, Vol. 19, No. 11, pp. 1236—1250, Nov. 1997.

[36] P. Ishwar and P. Moulin, "On the Equivalence of Set-Theoretic and Maxent MAP Estimation," *IEEE Trans. on Signal Processing*, Vol. 51, No. 3, pp. 698—713, March 2003.

[37] I. Csiszár, "Maxent, Mathematics, and Information Theory," in *Maximum Entropy and Bayesian Methods*, K. M. Hanson and R. N. Silver (eds.), pp. 35—50, Kluwer, 1996.

[38] I. Csiszár, "I-Divergence Geometry of Probability Distributions and Minimization Problems," *Annals of Probability*, Vol. 3, pp. 146—158, 1975.

[39] I. Csiszár and G. Tusnady, "Information Geometry and Alternating Minimization Procedures," *Statistics and Decisions*, Vol. 1, supplement issue, pp. 205—237, 1984.

[40] A. Gunawardana and W. Byrne, "Convergence Theorems for Generalized Alternating Minimization Procedures," *J. of Machine Learning Research*, Vol. 6, pp. 2049—2073, 2005.

[41] P. Perona and J. Malik, "Scale-Space and Edge Detection Using Anisotropic Diffusion," *IEEE Trans. on PAMI*, Vol. 12, No. 7, pp. 629—639, July 1990.

[42] L. Alvarez, P.-L. Lions and J.-M. Morel, "Image Selective Smoothing and Edge Detection by Nonlinear Diffusion. II," *SIAM J. Numer. Anal.*, Vol. 29, No. 3, pp. 845—866, June 1992.

[43] G. Sapiro, "From Active Contours to Anisotropic Diffusion: Connections Between Basic PDE's in Image Processing," *Proc. ICIP*, pp. I. 477—480, Lausanne, Switzerland, 1996.

[44] J. Koenderink, "The Structure of Images," *Biol. Cybern.*, Vol. 50, pp. 363—370, 1984.

[45] M. Bertalmio, G. Sapiro, V. Caselles and C. Ballester, "Image Inpainting", *ACM SIGGRAPH*, pp. 417—424, New Orleans, 2000.

[46] S. Baker and T. Kanade, "Limits on Superresolution and How to Break Them," *IEEE Trans. PAMI*, Vol. 24, No. 9, pp. 1167—1183, 2002.

[47] D. Robinson and P. Milanfar, "Statistical Performance Analysis of Super-Resolution," *IEEE Trans. Image Processing*, Vol. 15, No. 6, pp. 1413—1428, June 2006.

[48] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan-Kaufman, 1988.

[49] L. E. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, Vo. 72, No. 2, pp. 257—286, 1989.

[50] J. Yedidia and W. T. Freeman, "Understanding Belief Propagation and its Generalizations," *MERL Tech. Rep.*, 2001.

[51] M. J. Wainwright and M. I. Jordan, "Graphical Models, Exponential Families, and Variational Inference" *Tech. Rep. 649* (103 pages), Dept. of Statistics, UC Berkeley, 2003.

[52] M. J. Wainwright and M. I. Jordan, "A Variational Principle for Graphical Models," Chapter 11 in *New Directions in Statistical Signal Processing*, S. Haykin *et al.*, Eds., MIT Press, 2005.

[53] S. Roth and M. J. Black, "Fields of Experts: A Framework for Learning Image Priors," *Proc. CVPR*, 2005.

[54] P. Moulin and J. Liu, "Statistical Imaging and Complexity Regularization," *IEEE Trans. Information Theory*, Special issue on information-theoretic imaging, Vol. 46, No. 5, pp. 1762—1777, Aug. 2000.

[55] J. Liu and P. Moulin, "Complexity-Regularized Image Denoising," *IEEE Trans. Image Processing*, Vol. 10, No. 6, pp. 841—851, June 2001.

[56] W. T. Freeman, E. C. Pasztor and O. T. Carmichael, "Learning Low-Level Vision," *Int. J. Comp. Vision*, 2000.