IMAGE RESTORATION

ZHENGWEI SHEN
DEPARTMENT OF MATHEMATICS OF UNIVERSITY SCIENCE AND TECHONOLGY, BEIJING, CHINA

1. Mathematical representation of images in a function space

Since the image function is a mathematical object, thus, we can treat it as such and apply mathematical operations to it. These mathematical operations are summarized by the term *image processing techniques*, and range from *statistical methods*, *morphological operations*, *variational methods* and *PDE*.

While image processing is indeed concerned with digital images, the methods used are often motivated from considerations in the continuous case, that is methods are formulated for the analog image. Let $u(x, y)$ be a continuous image defined on a rectangular domain $\Omega = (a, b) \times (c, d) \subset \mathbb{R}^2$. Thus, the grey value image can be defined by $u : \Omega \to \mathbb{R}$.

Image can be treated as piecewise constant function.

2. What is image restoration?

Suppose that $f(x, y)$ is the true image that we would like to recover from the degraded measurement $g(x, y)$, where (x, y) are spatial coordinates. For a linear shift-invariant system, the imaging process can be formulated as

$$g(x, y) = h(x, y) * f(x, y) + n(x, y) \tag{1}$$

where $*$ is the convolution operation, $h(x, y)$ is the PSF of the imaging system, and $n(x, y)$ is the additive white Gaussian noise(AWGN). See Fig. (1).

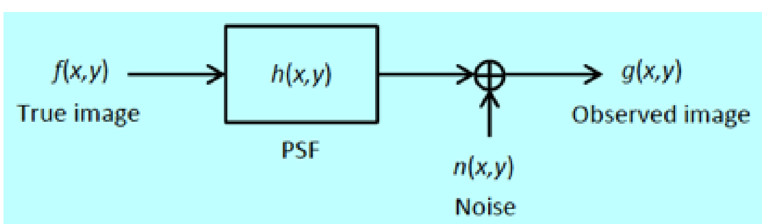

Figure 1. Linear shift-invariant image degradation model

Date: Sep. 12, 2017.

FIGURE 2. Image denoising

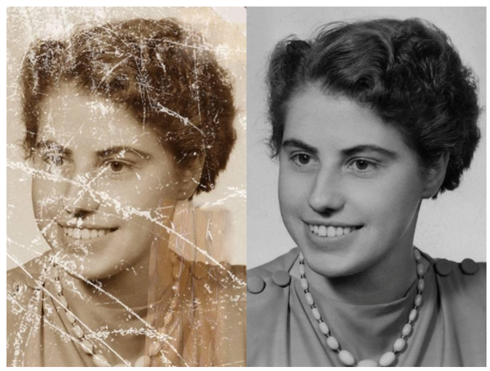

FIGURE 3. Image impainting

FIGURE 4. Image deblurring

The imaging formulation can also be done in matrix-vector form or in frequency domain. Defining $\mathbf{g}$, $\mathbf{f}$, and η as the vectorized versions of $g(x;y)$, $f(x;y)$, and $n(x;y)$, respectively, the matrix-vector formulation is

$$\mathbf{g} = \mathbf{H}\mathbf{f} + \eta \tag{2}$$

where $\mathbf{H}$ is a two-dimensional sparse matrix w.r.t. different boundary conditions. For instance, with periodic extension, matrix $\mathbf{H}$ is Block Circulant with Circulant Blocks(BCCB), which can be diagonalized by discrete Fourier transform.

Different linear-invariant operator $\mathbf{H}$ corresponding different image restoration task, e.g. an identity kernel I for denoising, blur kernel for deblurring and a restriction operator for inpainting, partial Random transform for CT imaging, partial Fourier transform for MR images. In essence, it is an inverse problem. However, many inverse problems that come from engineering or physics usually formulated as model (2). However, this inverse problem is ill-posed which meaning that the solution does not satisfy at least one of the following: existence, uniqueness, or stability. Regularization techniques are often adopted to obtain a solution with desired properties, indicating a knowledge of prior information about the true image.

On the other hand, the Fourier-domain version of the imaging model is

$$G(u, v) = H(u, v)F(u, v) + \eta(u, v) \tag{3}$$

FIGURE 5. Image super-resolution

3. IMAGE RESTORATION METHODS

As George Box once said "All models are wrong; but some are useful". Image models underlying all existing image restoration algorithms, no matter explicitly or implicitly stated, can be classified into two categories: deterministic and statistical. Deterministic models include those studied in functional analysis (e.g., Sobolov and Besov-space functions) and partial differential equations (PDE); statistical models include Markov Random

Field (MRF), conditional random field (CRF), Gaussian scalar mixture (GSM) and so on. Despite the apparent difference at the surface, deterministic and statistical models have intrinsic connections

3.1. **Least Squares Estimation.**

3.1.1. *Matrix norm and singular value.* For matrix H, the induced 2-norm of H is defined a s

$$\|H\|_2 = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{\|x\|_2 = 1} \|Ax\|_2 \tag{4}$$

Thus, the induced norm measures the amount of "amplification" the matrix H provides to vectors on the unit sphere in C^n, i.e. it measures the "gain" of the matrix.

Some facts of matrix

- if matrix $A = A^*$(i.e. A equals its Hermitian transpose, in which case we say A is Hermitian), then there exists a unitary matrix U (i.e. $U^*U = I$), such that

$$A = U\Sigma U^* \tag{5}$$

  where Σ is a diagonal matrix, and the diagonal entry are the eigenvalues of A.
- For any matrix B, both B^*B and BB^* are Hermitian, thus can be diagonalized by unitary matrices;
- For any matrix B, the eigenvalues of B^*B and BB^* are always real and non-negative, since B^*B and BB^* are positive-semidefinite matrix.

The singular value decomposition(SVD) of $H \in C^{m\times n}$

$$H = U\Sigma V^* \tag{6}$$

where $U \in C^{m\times m}$, $V \in C^{n\times n}$, and $\Sigma \in C^{m\times n}$ and $U^*U = I$, $V^*V = I$. And the "diagonalize" entry of Σ, $\sigma_i = \sqrt{\lambda_i(H^*H)}$, with $0 \le i \le rank(A)$. σ_i are termed the singular values of A, and are arranged in order of descending magnitude, i.e.

$$\sigma_1 \ge \sigma_2 \ge .... \ge \sigma_r \ge 0$$

Also, the $\|H\|_2$(a.k.s, the spectral norm) can be defined as the largest singular value(non-negative real numbers) of H, i.e. the square root of the largest eigenvalue of the positive-semidefinite matrix H^TH, where notice that matrix H^*H is symmetric matrix.

$$\|H\|_2 = \sqrt{\lambda_{max}(H^*H)} = \sigma_{max}(H) \tag{7}$$

where H^* denotes the conjugate transpose of H.

Proof to this proposition, please recall the (4) and the SVD decomposition of H.

3.1.2. *Least Squares Estimation.* The least squares estimation of $\mathbf{f}$ from the observation $\mathbf{g{=}HF} + \eta$ is to minimize the sum of squared differences between the real observation $\mathbf{f}$ and the observation $\mathbf{g}$. Thus, the *cost function* to be minimized can be written as

$$\sum\sum_{(i,j)} |f[i,j] - h[i,j] * f[i,j]|^2$$

in spatial domain or in matrix-vector notation, the solution can be defined as follows

$$\mathbf{f}^* = argmin_{\mathbf{f}} \|g - Hf\|^2 = H^+ g \tag{8}$$

where H^+ is known as the pseudo-inverse of H, which can be derived by setting the derivative of the cost function to zero

$$\frac{\partial}{\partial \mathbf{f}} \|g - Hf\|^2 = 2H^T Hf - 2H^T g = 0$$

Consider the SVD of $H_{m\times n} = U_{m\times m}\Sigma_{m\times n}V^T_{n\times n}$, with Σ is an diagonal matrix with singular values on the diagonal. The pedudo-inverse H^+ can be shown to be equal to

$$H^+ = V\Sigma^{-1}U^T$$

Then

$$\mathbf{f}^* = V\Sigma^{-1}U^T g = \sum_{i=1}^{r} \frac{1}{\sigma_i}(u_i^T g)v_i$$

As seen in this equation, a system with small singular values σ_i is not stable, since a small perturbation in g would lead to large change in the solution $\mathbf{f}^*$

To estimate $\mathbf{f}$ from the observation $\mathbf{g{=}HF} + \eta$ is typically ill-posed problem.

3.1.3. *Least Squares Estimation with regularization.* An alternative way to improve the conditioning of the system and also to give preference to a solution with desirable properties is to incorporate regularization terms into the problem.

Tikhonov regularization is a commonly used technique in LSE: the cost function to be minimized is modified to

$$f^* = argmin_f C(f) = \|g - Hf\|^2 + \lambda \|Df\|^2 \tag{9}$$

where λ is a nonnegative regularization parameter controlling the trade-off between the fidelity and regularization terms, and D is an identity matrix or a high-pass filter matrix used to impose smoothness on the solution, for instance *gradient operator*, *Laplacian operator*, *Wavelet operator......*

$$D = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & & \\ & & & -1 & 1 \end{pmatrix} \tag{10}$$

When taking $D = I$ be an identity matrix, we get the *standard* Tikhonov regularization formulation

$$f^* = argmin_f C(f) = \|g - Hf\|^2 + \lambda\|f\|^2 \tag{11}$$

If the regularization parameter λ is chosen too small, (9) is too close to the original problem and instabilities have to be expected. If λ is chosen too large, the problem we solve has only little connection with the original problem. Thus, finding the optimal parameter λ is a tough problem.

On the other hand, Tikhonov regularization has an important equivalent formulation as

$$f^* = argmin_f \|g - Hf\|^2 \quad subject \quad to \quad \|Df\|^2 \le \varepsilon \tag{12}$$

where ε is a positive constant.

Thus, optimization (12) is a linear LSE problem with a quadratic constraint, and using the Lagrange multiplier formulation. Then,

$$L(f, \lambda) = \|g - Hf\|_2^2 + \lambda(\|Df\|_2^2 - \varepsilon)$$

It can be shown that if $\varepsilon \le \|f_{LS}\|$, where f_{LS} denotes the Least square solution of $\min_f \|g - Hf\|_2^2$, then the solution of (12) is identical to the solution (9) for an appropriately chosen λ, and there is a monotonic relation between the parameters ε and λ.

3.2. **Solutions to Least square estimation.** Some notions: Suppose x is a column vector, A is a matrix, then we have

- $$\frac{\partial(Ax)}{\partial x} = A \tag{13}$$
- $$\frac{\partial(x^T A)}{\partial x^T} = A \tag{14}$$
- $$\frac{\partial(x^T A)}{\partial x} = A^T \tag{15}$$
- $$\frac{\partial(x^T A x)}{\partial x} = x^T(A^T + A) \tag{16}$$

With Tikhonov regularization, the solution (in matrix-vector notation) is

$$\mathbf{f}^* = (H^T H + \lambda D^T D)^{-1} H^T g \tag{17}$$

By taking $D = I$,

$$\mathbf{f}^* = (H^T H + \lambda I)^{-1} H^T g \tag{18}$$

How to solve this problem ? In fact, this direct solution in matrix-vector form is not computationally feasible because of the dimensions involved in computing the inverse of $(H^T H + \lambda D^T D)^{-1}$ or $(H^T H + \lambda I)^{-1}$; instead, iterative matrix inversion methods could be used.

On the other hand, defining $D(u,v)$ as the Fourier domain version of D, the direct solution in Fourier domain is

$$F(\hat{u,}v) = \frac{H^*(u,v)}{|H(u,v)|^2 + \lambda |D(u,v)|^2} G(u,v) \tag{19}$$

or

$$F(\hat{u,}v) = \frac{H^*(u,v)}{|H(u,v)|^2 + \lambda I} G(u,v) \tag{20}$$

which can be computed very efficiently using fast Fourier transform.

Actually, this solution extends the inverse filter solution, which just is the least square estimation solution

$$F(\hat{u,}v) = \frac{G(u,v)}{H(u,v)}$$

and with proper choices λ and $D(u,v)$, it does not suffer from noise amplification when $H(u,v)$ is zero or close to zero.

3.2.1. *Solutions in Frequency domain.* If the image restoration model is written as the matrix-vector formulation

$$\mathbf{g} = \mathbf{Hf} + \eta \tag{21}$$

Then, with periodic extension, matrix $\mathbf{H}$ is Block Circulant with Circulant Blocks(BCCB), which can be diagonalized by discrete Fourier transform.

For example, consider a 3×3 image X and let h is a 3×3 point spread function(PSF). Let the output image B is a 3×3 image.

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{pmatrix} \tag{22}$$

$$P = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix} \tag{23}$$

$$B = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} \tag{24}$$

Then we have the following BCCB matrix

$$\begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \\ b_{12} \\ b_{22} \\ b_{32} \\ b_{13} \\ b_{23} \\ b_{33} \end{bmatrix} = \left[\begin{array}{ccc|ccc|ccc} p_{22} & p_{12} & p_{32} & p_{21} & p_{11} & p_{31} & p_{23} & p_{13} & p_{33} \\ p_{32} & p_{22} & p_{12} & p_{31} & p_{21} & p_{11} & p_{33} & p_{23} & p_{13} \\ p_{12} & p_{32} & p_{22} & p_{11} & p_{31} & p_{21} & p_{13} & p_{33} & p_{23} \\ \hline p_{23} & p_{13} & p_{33} & p_{22} & p_{12} & p_{32} & p_{21} & p_{11} & p_{31} \\ p_{33} & p_{23} & p_{13} & p_{32} & p_{22} & p_{12} & p_{31} & p_{21} & p_{11} \\ p_{13} & p_{33} & p_{23} & p_{12} & p_{32} & p_{22} & p_{11} & p_{31} & p_{21} \\ \hline p_{21} & p_{11} & p_{31} & p_{23} & p_{13} & p_{33} & p_{22} & p_{12} & p_{32} \\ p_{31} & p_{21} & p_{11} & p_{33} & p_{23} & p_{13} & p_{32} & p_{22} & p_{12} \\ p_{11} & p_{31} & p_{21} & p_{13} & p_{33} & p_{23} & p_{12} & p_{32} & p_{22} \end{array}\right] \begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \\ x_{12} \\ x_{22} \\ x_{32} \\ x_{13} \\ x_{23} \\ x_{33} \end{bmatrix}$$

FIGURE 6. Matrix-vector multiplication formulation with BCCB matrix

For the regularization Tikohonov solution in the Frequency domain

$$F(\hat{u}, v) = \frac{H^*(u,v)}{|H(u,v)|^2 + \lambda I} G(u,v) \tag{25}$$

can be implemented using the following Matlab

$$h = fspecial(masktype, masksize, maskvariation);$$

$$bluredim = imfilter(cleanimage, h,' circular');$$

Also, you can add the noise into the bluredim in this step

$$H = psf2otf(h, [m, n]);$$

where $m \times n$ is the size of the image;

$$HTH = conj(H). * H$$

$$HTg = conj(H). * fft2(bluredim)$$

$$f = ifft2(HTg./(HTH + \lambda))$$

3.3. **Iterative solution.** In cases where there is no direct solution or a direct solution is not computationally feasible, an iterative scheme is adopted. The steepest descent method(SDM) updates an initial estimate iteratively in the reverse direction of the gradient of the cost function $C(f)$ in Eq.(9).

An iteration of the SDM is

$$\begin{aligned} f^{n+1} &= f^n - \alpha \frac{\partial C(f)}{\partial f}|_{f^n} \\ &= f^n - \alpha \frac{\partial C(f)}{\partial f}|_{f^n} \end{aligned} \tag{26}$$

When there is no regularization term, i.e. $C(f) = \frac{1}{2}\|g - Hf\|_2^2$, then the above iteration is

$$f^0 = g$$
$$f^{n+1} = f^n - \alpha \frac{\partial C(f)}{\partial f}|_{f^n} = f^n + \alpha H^T(g - Hf^n) = \alpha H^T g + (I - \alpha H^T H) f^n \tag{27}$$

which is also known as the Landweber iteration.

On the other hand, the iteration algorithm to the regularization case, i.e. $C(f) = \frac{1}{2}\|g - Hf\|_2^2 + \lambda \|Df\|_2^2$,

(28)
$$f^0 = g$$
$$f^{n+1} = f^n - \alpha \frac{\partial C(f)}{\partial f}|_{f^n} = f^n + \alpha(H^T(g - Hf^n) + \lambda D^T D f^n) = f^n + \alpha(H^T g - (H^T H + \lambda D^T D) f^n))$$

Notice that in the implementation, you can take $D = I$.

4. Maximum A Posterior point view to the image restoration

Consider the image restoration problem

$$y = k * x + \eta$$

The Bayesian law of image restoration is to maximize the following conditional probability density relationship (Maximum A Posterior)

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

where $p(x|y)$ is the *posterior* that is the density of x given the observation y, $p(y|x)$ is the *likelihood* that is the density given x, $p(x)$ is the *prior* of x, since y is known, then $p(y)$ is a constant.

That is

$$\begin{aligned} x^* = argmax_x p(x|y) &= argmax_x \underbrace{p(y|x)}_{likelihood} \underbrace{p(x)}_{prior} \\ &= argmax_x \{\log p(y|x) + \log p(x)\} \\ &= argmin_x \{-\log p(y|x) - \log p(x)\} \end{aligned} \tag{29}$$

The likelihood function is governed by physical considerations concerned with the data-acquisition device. Assuming the noise η is Gaussian white noise with σ_η^2 variance, then

$$p(y|x) = \frac{1}{\sqrt{2\pi\sigma_\eta^2}} exp(-\frac{(y - k * x)^2}{2\sigma_\eta^2})$$

Some other assumption to noise including Laplacian noise, Poisson noise and Multiplicative noise.

How about the prior of $p(x)$? To get the conditional distribution of the parameters x given the data y, we need the distribution of the parameters x in the absence of any data. This is called the prior.

It is often to consider its coefficients distribution under some sparse representative dictionary $\alpha = Dx$. Empirically, the probability distribution of α to be *peaked at zero* and have *high kurtosis*, also with *heavy-tailed*, such as Laplace distribution and generalized Gaussian distribution.

Generally, the convenient parameterization of the prior distribution of α is supposed to be Gibbsian form

$$p(\alpha) = \frac{1}{C} exp(-\frac{1}{\beta} T(\alpha)) \tag{30}$$

where $T(\alpha)$ is a prior energy function determining the shape of the prior distribution.

- The Laplace distribution can be formulated as

$$p(\alpha) = \frac{1}{2b} exp(-\frac{|\alpha|}{b}) \tag{31}$$

  corresponds to the ℓ_1-norm(TV) regularization term in MAP formulation.
- The Cauchy distribution

$$p(\alpha) = \frac{1}{2b} exp(-\frac{\log(1+\alpha^2)}{b}) \tag{32}$$

- The Generalized Gaussian distribution.

$$p(\alpha) = \frac{q}{2\Gamma(1/q)} e^{-\frac{|\alpha|^p}{b}} \tag{33}$$

  where $b > 0$ and $p > 0$ are appropriate choices of the parameters.
- And the Student t distribution with $v > 0$

$$p(\alpha) = (1 + \frac{1}{b}\frac{\alpha^2}{v})^{-\frac{v+1}{2}} \tag{34}$$

If we take the noise η is Gaussian white noise

$$p(y|x) = \frac{1}{\sqrt{2\pi\sigma_\eta^2}} exp(-\frac{(y-k*x)^2}{2\sigma_\eta^2})$$

Then

$$-\log p(y|x) = \frac{(y-k*x)^2}{2\sigma_\eta^2}$$

Thus the fidelity term

$$\frac{(y-k*x)^2}{2\sigma_\eta^2}$$

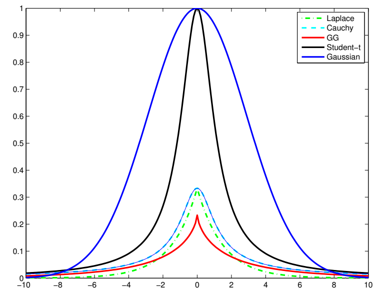

FIGURE 7. The pdf Gaussian, Generalized Gaussian, Laplace, Cauchy, Student t distribution with the mean is 0.

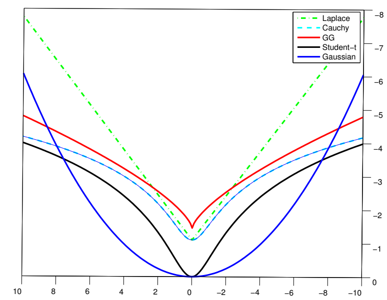

FIGURE 8. The log of these pdf. Notice the not log-convex property of Generalized Gaussian, Cauchy, and Student t distribution

If we take the Laplace distribution of transform coefficients α , i.e.

$$p(\alpha) = \frac{1}{2b} exp(-\frac{|\alpha|}{b})$$

corresponding to the

$$-\log p(\alpha) = \frac{|\alpha|}{b}$$

Then image restoration model is

$$\min_x \frac{(y-k*x)^2}{2\sigma_\eta^2} + \frac{|\alpha|}{b}$$

That is

$$\min_x \frac{1}{2}|y-k*x|^2 + \frac{\sigma_\eta^2}{b}|\alpha| \tag{35}$$

Cauchy distribution of α.

$$\min_x \frac{1}{2}|y-k*x|^2 + \frac{\sigma_\eta^2}{b}\log(1+\alpha^2) \tag{36}$$

Generalized Gaussian distribution of α.

$$\min_x \frac{1}{2}|y-k*x|^2 + \frac{\sigma_\eta^2}{b}|\alpha|^p \tag{37}$$

The challenge of image restoration and the methods to make it

- The challenge is how to recovery more details or textures of the images from the observed image y. i.e. to improve the PNSR.

- Recall the image restoration model

$$\min_x \|y - k * x\|^2 + \lambda \phi(Dx) \tag{38}$$

- We can summerize the existing works to do it as follows:
  - Adaptive selection of regularization λ;
  - Accurate selection of regularization term $\phi(\cdot)$;
  - "Good" sparse representation of image x;
  - Multi-hiberarchy decomposition of the image x.
  - Algorithm design.

5. Total variation(TV) image restoration

TV is a powerful notion in robust statistics and robust signal estimation. Consider a set of independent observations $\{x_1, x_2, ..., x_N\}$ generated from an underlying scalar x(due to noise or random perturbations).

The question is how to solve(estimate) the underlying scalar x from the observations $\{x_1, x_2, ..., x_N\}$?.

Consider the Mean Squared Error estimation function:

$$x^* = argmin_x e_2(x) = \frac{1}{N}\sum_{k=1}^{N}(x - x_k)^2 \tag{39}$$

The best solution/estimator $\hat{x}^{(2)}$to (39) is precisely the mean of the samples:

$$x^* = \frac{1}{N}\sum_{k=1}^{N} x_k \tag{40}$$

If the noise or random perturbations is Gaussian distribution, then the Mean Squared Error estimation is appropriate and is precisely the Maximum Likelihood(ML) estimator.

For the linear measurement model

$$y_i = a_i^T x + \mu_i, i = 1, ...m$$

where $x \in R^n$ is a vector of unknown parameters, μ_i is the i.i.d measurement noise with density $p(z)$

Then, the MLE of x is to solve

$$maximize_x E(x) = \sum_{i=1}^{m} \log p(y_i - a_i^T x)$$

If the noise density

$$p(z) \propto \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{z^2}{2\sigma^2}}$$

Then,

$$maximize_x E(x) = \sum_{i=1}^{m} \log p(y_i - a_i^T x)$$

is equivalent to

$$minimize\|y - Ax\|^2$$

which is a least-squares approximation.

Consider another case, i.e. if there exists any spontaneous deviation observations in $\{x_1, x_2, ..., x_N\}$, for example, suppose for a set of 10 observations, the mean estimator is $\bar{x} = 10$; but now, another experimenter obtains a very similar set of 10 observations: all data are identical except for the last datum $x_{10} = 100$. Then the Least square estimator undergoes a substantial change from 10 to 20, which indicates the Mean is not a robust estimator—any erratic change of the data can cause a big change in the estimator.

5.1. **Total variation.** How to improve the robustness of the estimation error $e_2(x)$, i.e. the 2-norm estimator?

A slight modification to $e_2(x)$ is:

$$e_1(x) = argmin_x \sum_{k=1}^{N} |x - x_k| \tag{41}$$

The associated optimal estimator to (41) is the Median of the samples.

$$\hat{x}^{(1)} = median\{x_k\} = x_{N/2} \tag{42}$$

The change in the optimal estimator is bounded by the mutual spacing of the majority, and is thus insensitive to any single trouble maker.

Definition 1. *The Total Variation of a differentiable function f, defined on an interval $[a,b] \subset \mathcal{R}$, has the following expression if f' is Riemann integrable*

$$V_b^a(f) = \int_a^b |f'(x)| dx \tag{43}$$

For the $f \in \mathcal{R}^n$ and differentiable defined on a bounded open set $\Omega \subset \mathcal{R}^n$, TV of f is defined as

$$V_\Omega(f) = \int_\Omega |\nabla f(x)| dx \tag{44}$$

where $|\nabla f(x)|$ denote the magnitude of the gradient of $f(x)$, i.e. $|\nabla f(x)| = \sqrt{(\nabla_x f)^2 + (\nabla_y f)^2}$
.

For a real-valued function f, defined on an interval $[a,b] \subset \mathcal{R}$, TV is defined as the following quantity

$$V_b^a(f) = \sup_{\mathcal{P}} \sum_{i=0}^{N} |f(x_{i+1}) - f(x_i)| \tag{45}$$

where the Supremum runs over the set of all partitions $\mathcal{P} = \{x_0, ...x_N | \mathcal{P}$ is a partition of $[a, b]\}$ of the given interval.

Definition 2. *A real-valued function f on the real line is said to be of Bounded variation(BV) function on a chosen interval $[a, b] \subset \mathcal{R}$ if its total variation is finite, i.e. $V_a^b < +\infty$*

$$f \in BV[a, b]$$

In particular, if $f \in BV[a, b]$ and f is monotone(no matter the shape of the function f) in $[a, b]$ and $f(a) = \alpha, f(b) = \beta$, then we have

$$TV(f) = \beta - \alpha$$

So, the Total Variation of a a signal measures how much the signal changes between signal values. The total variation of an N-point signal $x[n]$, $1 \le n \le N$ is defined as

$$TV(x) = \sum_{n=2}^{N} |x[n] - x[n-1]| \tag{46}$$

The TV of x can also be written as

$$TV(x) = \|Dx\|_1 \tag{47}$$

where $\|Dx\|_1$ is the ℓ_1 norm and

$$D = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & & \\ & & & -1 & 1 \end{pmatrix} \tag{48}$$

One questions is what kind of function f can get smaller TV? What result will you obtain if you try to minimize the TV of a function f?

5.2. **Total variation for two-dimensional images.** We now write the images in the continuous setting: as gray-level values functions $u(x, y) : \Omega \to \mathbb{R}$, where $\Omega \subset \mathbb{R}$ is the definition domain of (x, y).

The TV of it is denoted by

$$\int_\Omega |Du(x, y)| dxdy = \int_\Omega \sqrt{(\nabla_x u)^2 + (\nabla_y u)^2} dxdy$$

where $|Du(x, y)| = \sqrt{(\nabla_x u)^2 + (\nabla_y u)^2}$ is the magnitude of the gradient to $u(x, y)$.

This type TV is called isotropic total variation is proposed by Rudin, Osher and Fatemi, i.e. the ROF model

In addition to the isotropic models, a variation that is sometimes used, since it may sometimes be easier to minimize, is the anisotropic TV model for a qualitative improvement at corners are defined

$TV_{anisotropic}(u) = \int_\Omega |\nabla_x u| + |\nabla_y u| dxdy.$

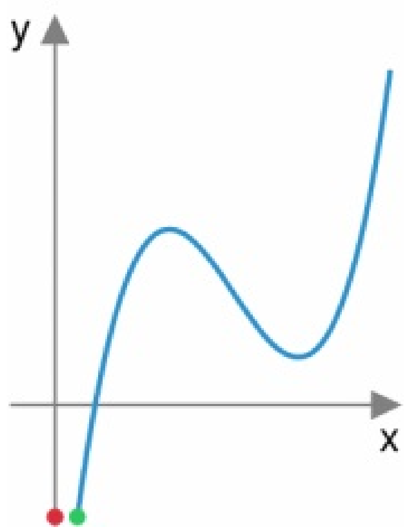

FIGURE 9. Total Variation

5.3. **ROF model.** We assume the observed image $u_0(x, y)$ is the signal $u(x, y)$ is corrupted by additive White Gaussian noise

$$u_0(x, y) = u(x, y) + \eta(x, y) \tag{49}$$

In fact, the TV denoising equation can be treated as a constrained optimization problem. i.e. Rudin, Osher and Fatemi(ROF)

$$min_u \int_\Omega |Du(x, y)|dxdy \tag{50}$$

subject to

$$\|u(x, y) - u_0(x, y)\|_2^2 \leq \sigma^2$$

Writing it as the non-constricted optimization problem as follows

$$\begin{aligned} u^* &= argmin_{u \in L^2(\Omega)} E(u(x, y)) = \underbrace{\int_\Omega (u_0(x, y) - u(x, y))^2 dxdy}_{Fadelity} + \lambda \underbrace{\int_\Omega |Du(x, y)|dxdy}_{TV} \\ &= \Phi(u) + \lambda\Psi(u) \end{aligned} \tag{51}$$

where we consider u in the space $L^2(\Omega)$ of functions which are square-integrable, since the energy will be infinite if u is not.

Remark 1. *This equation*(51) *can also be interpreted as the Variational method, i.e. to minimize an energy functional of which we will introduce in the next section.*

The QUESTION:
Why do not take

$$\int_\Omega (|Du|)^2 dxdy = \int_\Omega |\nabla_x u|^2 + |\nabla_y u|^2 dxdy$$

to be the penalty function, i.e. regularization term $\Psi(u)$?

As we have mentioned above, ℓ_2 norm is the Maximum Likehood estimation, which is easily sensitive to deviation observation(i.e. the singular point or the image edges). In fact, taking $\int_\Omega (|Du|)^2 dxdy$ to be the regularization term is equivalent to be the Tikohonov regularization, which is the linear filter that will smooth the edges of the solution. Since, the associated Euler-Lagrange Equations to

$$-\lambda \Delta u + u - u_0 = 0$$

On the other hand, ℓ_1 norm will penalized highly oscillatory solutions while allowing jumps(image edges) in the regularized solution. That is, the TV regularization results in signals that are approximately Piecewise Constant. The the associated Euler-Lagrange Equations to (51) is

$$-\lambda div(\frac{\nabla u}{|\nabla u|}) + u - u_0 = 0$$

Remark 2.
- *Minimizing $\int_\Omega |\nabla u|^2 dxdy$, the Euler-Lagrange Equations*

(52) $$\triangle u = 0$$

- *While, Minimizing $\int_\Omega |\nabla u| dxdy$, the Euler-Lagrange Equations*

(53) $$div(\frac{\nabla u}{|\nabla u|}) = 0$$

5.4. **Algorithms to the TV Denoising and TV deblurring.**

5.4.1. *Steepest-Descent method.* The minimization of this objective functions is complicated by the fact that the ℓ_1 norm is non-smooth(not differentiable). The Euler-Lagrange equation is

(54) $$0 = -div(\frac{\nabla u}{|\nabla u|}) + \lambda(u - u_0)$$

The Steepest-Descent Marching with artificial time t:

(55) $$\frac{\partial u}{\partial t} = div(\frac{\nabla u}{|\nabla u|}) - \lambda(u - u_0)$$

The nonlinear E-L equation (54) and (55) are both of Elliptic type but degenerate due to the gradient term in the denominators.

Algorithms to Eq.(55) with explicitly difference approximation method. We first denote the difference operator of the horizontal derivatives:

(1) Forward difference : $\nabla_x^+ = u_{i+1,j} - u_{i,j}$;
(2) Backward difference : $\nabla_x^- = u_{i,j} - u_{i-1,j}$;

The difference operator of the vertical derivatives:
(1) Forward difference : $\nabla_y^+ = u_{i,j+1} - u_{i,j}$;
(2) Backward difference : $\nabla_y^- = u_{i,j} - u_{i,j-1}$;

The Central difference operator of the horizontal derivatives:
$\frac{\nabla_x^+ + \nabla_x^-}{2} = \frac{(u_{i+1,j} - u_{i+1,j}) + (u_{i,j} - u_{i,j-1})}{2}$;
Note that Central difference are undesirable for TV discretization because they miss thin structures.

The Minmod difference operator of the horizontal derivatives:
$m(\nabla_x^+, \nabla_x^-) = (\frac{sign(\nabla_x^+) + sign(\nabla_x^-)}{2}) \min(|\nabla_x^+|, |\nabla_x^-|)$

$$
\begin{aligned}
&\frac{u_{i,j}^{n+1} - u_{i,j}^n}{\triangle t} = \\
&[\nabla_x^- (\frac{\nabla_x^+ u_{i,j}^n}{\sqrt{(\nabla_x^+ u_{i,j}^n)^2 + (m(\nabla_y^+ u_{i,j}^n, \nabla_y^- u_{i,j}^n))^2}}) \\
&+ \nabla_y^- (\frac{\nabla_y^+ u_{i,j}^n}{\sqrt{(\nabla_y^+ u_{i,j}^n)^2 + (m(\nabla_x^+ u_{i,j}^n, \nabla_x^- u_{i,j}^n))^2}})] \\
&- \lambda(u_{i,j}^n - u_{i,j}^0)
\end{aligned} \tag{56}
$$

where $i, j = 1, ...N-1, u_{0,j}^n = u_{1,j}^n, u_{N,j}^n = u_{N-1,j}^n, u_{i,0}^n = u_{i,1}^n, u_{i,N}^n = u_{i,N-1}^n; i, j = 0, ...N$.

Give the initial value $u_{i,j}^0$ at position (i, j), the iteration is thus till convergence:

$$
\begin{aligned}
u_{i,j}^{n+1} = u_{i,j}^n + \triangle t \times &[\nabla_x^- (\frac{\nabla_x^+ u_{i,j}^n}{\sqrt{(\nabla_x^+ u_{i,j}^n)^2 + (m(\nabla_y^+ u_{i,j}^n, \nabla_y^- u_{i,j}^n))^2}}) \\
&+ \nabla_y^- (\frac{\nabla_y^+ u_{i,j}^n}{\sqrt{(\nabla_y^+ u_{i,j}^n)^2 + (m(\nabla_x^+ u_{i,j}^n, \nabla_x^- u_{i,j}^n))^2}})] \\
&- \triangle t \times \lambda(u_{i,j}^n - u_{i,j}^0)
\end{aligned} \tag{57}
$$

where $i, j = 1, ...N-1, u_{0,j}^n = u_{1,j}^n, u_{N,j}^n = u_{N-1,j}^n, u_{i,0}^n = u_{i,1}^n, u_{i,N}^n = u_{i,N-1}^n; i, j = 0, ...N$ and $\triangle t$ is a small positive time step parameter.

Remark 3. *(1) The discretization is symmetric through a balance of forward and backward differences.*
(2) In the divisions, notice that the numerator is always smaller than the denominator. In the special case that denominator is zero(where u is locally constant), the quotient is evaluated as $0/0 = 0$.

(3) Actually, regularization parameter λ is critical to the stability of convergence, and should be a dynamic values not a constant.

5.4.2. *Splitting Iterative shrinkage/thresholding algorithms and ADMM.* In the discrete setting, we often writhe the models as follows

$$\min_u \|u_0 - u\|_2^2 + \lambda \|Du\|_1 \tag{58}$$

If the we have

$$u_0 = Hu + \eta$$

then the TV model to recover u from u_0 will be denoted by

$$\min_u \|u_0 - Hu\|_2^2 + \lambda \|Du\|_1 \tag{59}$$

where the gradient operator D can be replaced by any other sparse representation operator, for instance, wavelet ......

The model (59) is not only used for denoising, but also can be used for debluring.

Now, we introduce one algorithm for solving (59), i.e. the alternating direction multipliers method(ADMM)

We firstly introduce an auxiliary variables $y = Du$ to (59). Then the unconstrained minimization problem (59) transformed to be an equivalent constrained problem of the form

$$\min_u \frac{1}{2}\|u_0 - Hu\|_2^2 + \lambda \|y\|_1 \quad s.t. \quad y = Du \tag{60}$$

We note that in (60), the objective function is seperable and the constraints are linear. Then the classical quadratic penalty method is applied to (60), which gives the following unconstrained formulation

$$\min_{u,y} \frac{1}{2}\|u_0 - Hu\|_2^2 + (\lambda \|y\|_1 + \frac{\beta}{2}\|y - Du\|_2^2) \tag{61}$$

where $\beta >> 0$ is a penalty parameter.

An alternating minimization algorithm with respect to u and y can be applied to solve (61). However, from the optimization theory, the solution of (61) well approximate that of (59) only when β becomes large, which results to numerical difficulties.

To avoid β going to infinity, the classical augmented Lagrangian method(ALM) is applied to solve (61), which the augmented Lagrangian function of (60) is formulated as

$$L_\beta(u, y, \mu) = \frac{1}{2}\|u_0 - Hu\|_2^2 + \lambda \|y\|_1 + \mu^T(y - Du) + \frac{\beta}{2}\|y - Du\|_2^2 \tag{62}$$

where μ is the Lagrange multiplier, β is the penalty parameter and L_0 will be the standard Lagrangian.

We claim that solving minimization problem (59) is equivalent to minimize the (62), i.e.

$$\min_{u,y,\mu} L_\beta(u,y,\mu) = \frac{1}{2}\|u_0 - Hu\|_2^2 + \lambda\|y\|_1 + \mu^T(y - Du) + \frac{\beta}{2}\|y - Du\|_2^2 \tag{63}$$

The ADMM algorithm to (63) is reformulated as follows

$$\begin{cases} u_{k+1} := argmin_u L_\beta(u, y_k, \mu_k) \\ y_{k+1} := argmin_y L_\beta(u_k, y, \mu_k) \\ \mu_{k+1} := argmin_\mu L_\beta(u_k, y_k, \mu) \end{cases} \tag{64}$$

The first subproblem

$$\begin{aligned} u_{k+1} : &= argmin_u \frac{1}{2}\|u_0 - Hu\|_2^2 + \mu_k^T(y_k - Du) + \frac{\beta}{2}\|y_k - Du\|_2^2 \\ &= argmin_u \frac{1}{2}\|u_0 - Hu\|_2^2 + \mu_k^T(y_k - Du) + \frac{\beta}{2}\|y_k - Du\|_2^2 + \frac{1}{2\beta}\|\mu_k\|_2^2 \\ &= argmin_u \frac{1}{2}\|u_0 - Hu\|_2^2 + \frac{\beta}{2}\|y_k - Du + \frac{1}{\beta}\mu_k\|_2^2 \end{aligned} \tag{65}$$

In the same way, the second subproblem

$$\begin{aligned} y_{k+1} : &= argmin_y \lambda\|y\|_1 + \mu_k^T(y - Du_k) + \frac{\beta}{2}\|y - Du_k\|_2^2 \\ &= argmin_y \lambda\|y\|_1 + \mu_k^T(y - Du_k) + \frac{\beta}{2}\|y - Du_k\|_2^2 + \frac{1}{2\beta}\|\mu_k\|_2^2 \\ &= argmin_y \frac{2\lambda}{\beta}\|y\|_1 + \|y - Du_k + \frac{1}{\beta}\mu_k\|_2^2 \end{aligned} \tag{66}$$

The third subproblem solved by steepest descent method

$$\mu_{k+1} := \mu_k + \delta(y_k - Dx_k) \tag{67}$$

where δ is the iterative step.
Solutions to these subproblems are summarized as

$$\begin{cases} u_{k+1} = (H^T H + \beta D^T D)^{-1}(H^T u_0 + \beta D^T(y_k + \frac{\mu_k}{\beta})) \\ y_{k+1} := Soft_{\frac{2\lambda}{\beta}}(Du_k - \frac{1}{\beta}\mu_k) \\ \mu_{k+1} := \mu_k + \delta(y_k - Dx_k) \end{cases} \tag{68}$$

where soft thresholding function is

$$Soft_\tau(x) = \begin{cases} x - \tau, & if \quad |x| > \tau \\ 0, & otherwise \end{cases} \tag{69}$$

6. The ideas of variational methods

6.1. **Calculus of variations.** In 1696, Johann Bernoulli proposed a variational problem:

"Given two points A and B in a vertical plane, what is the curve traced out by a point acted on only by gravity, which starts at A and reaches B in the shortest time"

The problem was generalized and an analytic method was given by Euler(1744) and Lagrange(1760).

Variational methods refer to the technique of optimizing the maximum or minimum of an *integral* involving unknown functions, which is the *functional*,e.g.

$$J(u) = \int_{x_1}^{x_2} F(x, u(x), u_x)dx$$

where u is the unknown functions.

A functional maps *functions* to scalars, so functionals have been described as *function of functions.*

The *calculus of variations* is concerned with the maxima or minima(collectively called extrema) of functionals. Finding the extrema of functionals is similar to finding the maxima or minima of functions. Thus, the extrema of functionals may be obtained by finding functions where the *functional derivatives* is equal to zero, i.e.

$$\frac{\partial J(u)}{\partial u} = 0 \tag{70}$$

This leads to solving the associated *Euler-Lagrange equation*

Consider the functional

$$J(u) = \int_{x_1}^{x_2} F(x, u(x), u_x)dx \tag{71}$$

where x_1 and x_2 are constants; $u(x)$ is twice continuously differentiable; $u_x = du/dx$; $F(x, u(x), u_x)$ is twice continuously differentiable w.r.t. its arguments $x, u(x), u_x$.

If functional $J(u)$ attains a local minimum at f, and let $\eta(x)$ be an arbitrary function that has at least one derivative and vanishes at the endpoints x_1 and x_2, i.e. $\eta(x_1) = \eta(x_2) = 0$, then for any number $\varepsilon \to 0$, we have

$$J(f) \leq J(f + \varepsilon\eta)$$

The term $\varepsilon\eta$ is called the variation of the function f and is denoted by δf. Substituting $f + \varepsilon\eta$ for u in the functional, then the result is a function of ε, i.e.

$$\Phi(\varepsilon) = J(f + \varepsilon\eta)$$

Since the functional $J(u)$ has a minimum for $u = f$, then $\Phi(\varepsilon)$ has a minimum at $\varepsilon = 0$ and thus,

$$\begin{aligned}\Phi'(0) &= \frac{d\Phi}{d\varepsilon}|_{\varepsilon=0} \\ &= \frac{dJ(f+\varepsilon\eta)}{d\varepsilon}|_{\varepsilon=0} \\ &= \int_{x_1}^{x_2} \frac{dF}{d\varepsilon}|_{\varepsilon=0} dx = 0\end{aligned} \tag{72}$$

Taking the total derivative of $F(x, u, u')$, where $u = f + \varepsilon\eta$ and $u' = f' + \varepsilon\eta'$ are functions of ε but x is not.

$$\frac{dF}{d\varepsilon} = \frac{\partial F}{\partial u}\frac{du}{d\varepsilon} + \frac{\partial F}{\partial u'}\frac{du'}{d\varepsilon} = \frac{\partial F}{\partial u}\eta + \frac{\partial F}{\partial u'}\eta'$$

Therefore,

$$\begin{aligned}\int_{x_1}^{x_2} \frac{dF}{d\varepsilon}|_{\varepsilon=0} dx &= \int_{x_1}^{x_2} \frac{\partial F}{\partial u}\eta + \frac{\partial F}{\partial u'}\eta' dx \\ &= \int_{x_1}^{x_2} \frac{\partial F}{\partial u}\eta - \eta\frac{d}{dx}\frac{\partial F}{\partial u'} dx + \frac{\partial F}{\partial u'}\eta|_{x_1}^{x_2} \\ &= 0\end{aligned} \tag{73}$$

where $F(x, u, u') \to F(x, f, f')$ when $\varepsilon \to 0$, then

$$\int_{x_1}^{x_2} \eta(\frac{\partial F}{\partial u} - \frac{d}{dx}\frac{\partial F}{\partial u'})dx = 0 \tag{74}$$

Then, we obtain the Euler-Lagrange equation to problem (71)

$$\frac{\partial F}{\partial u} - \frac{d}{dx}(\frac{\partial F}{\partial u_x}) = 0 \tag{75}$$

The left hand side of this equation is called the functional derivative of $J(u)$ and is denoted $\frac{\delta J}{\delta u(x)}$

For the two-dimensional case,

$$J(u) = \iint_\omega F(x, y, u(x, y), u_x, u_y)dxdy$$

The associated Euler-Lagrange equation is

$$\frac{\partial F}{\partial u} - \frac{d}{dx}(\frac{\partial F}{\partial u_x}) - \frac{d}{dy}(\frac{\partial F}{\partial u_y}) = 0 \tag{76}$$

Ex1: Let $J(u) = \iint_\Omega \rho(|\nabla u|)dxdy = \iint_\omega \rho[(u_x^2 + u_y^2)^{1/2}]dxdy$. Then the corresponding Euler-Lagrange equation is

$$
\begin{aligned}
0 &= \frac{d}{dx}(\frac{\partial F}{\partial u_x}) + \frac{d}{dy}(\frac{\partial F}{\partial u_y}) \\
&= \frac{d}{dx}[\rho'(|\nabla u|)\frac{u_x}{|\nabla u|}] + \frac{d}{dy}[\rho'(|\nabla u|)\frac{u_y}{|\nabla u|}] \\
&= div[\rho'(|\nabla u|)\frac{\nabla u}{|\nabla u|}]
\end{aligned}
\tag{77}
$$

In particular,

- Let $\rho(x) = x$. By (77), we have

$$div[\frac{\nabla u}{|\nabla u|}] = 0$$

- Let $\rho(x) = x^2$. By (77), we have

$$div[\nabla u] = \Delta u = 0$$

6.2. **Gradient descent flow method for solving functionals.** Generally, the Eular-Lagrange equation is an nonlinear PDE, it will be very difficult to compute directly after discretize it into a nonlinear algebra equations(it will be huge!). Consider the solution to Eular-Lagrange equation evolving with time, that is the solution of the PDE depends on time, i.e. $u(\cdot;t)$, and suppose that the energy functional $J(u(\cdot;t))$ will decrease with the time increasing. The gradient descent flow method to solve the variational problem.

For the 1D case, let $\varepsilon\eta$ be the change on function $u = f + \varepsilon\eta$, from $u(\cdot,t)$ to $u(\cdot,t+\Delta t)$. By (72),

$$
\begin{aligned}
&\frac{dJ(u)}{dt} = \frac{J(u(\cdot,t+\Delta t)) - J(u(\cdot,t))}{\Delta t} \\
&= \int_{x_1}^{x_2} \frac{dF}{dt}dx \\
&= \int_{x_1}^{x_2} \frac{\partial F}{\partial u}\frac{\partial u}{\partial t} + \frac{\partial F}{\partial u_x}u_{tx}dx \\
&= \int_{x_1}^{x_2} \frac{\partial F}{\partial u}\frac{\partial u}{\partial t}dx + \int_{x_1}^{x_2} \frac{\partial F}{\partial u_x}du_t \\
&= \int_{x_1}^{x_2} \frac{\partial F}{\partial u}\frac{\partial u}{\partial t}dx + u_t\frac{\partial F}{\partial u_x}|_{x_1}^{x_2} - \int_{x_1}^{x_2} \frac{\partial u}{\partial t}\frac{d}{dx}\frac{\partial F}{\partial u_x}dx \\
&= \int_{x_1}^{x_2} \frac{\partial u}{\partial t}(\frac{\partial F}{\partial u} - \frac{d}{dx}\frac{\partial F}{\partial u_x})dx
\end{aligned}
\tag{78}
$$

Suppose that $\frac{\partial u(x;t)}{\partial t}$ vanishing at endpoints x_1 and x_2, thus, we have

$$\begin{aligned} J(u(\cdot, t+\Delta t)) = J(u(\cdot, t)) + \\ \Delta t \int_{x_1}^{x_2} \frac{\partial u}{\partial t}(\frac{\partial F}{\partial u} - \frac{d}{dx}\frac{\partial F}{\partial u_x})dx \end{aligned} \tag{79}$$

From this equation, it can be easily found that, when taking

$$\frac{\partial u}{\partial t} = \frac{d}{dx}(\frac{\partial F}{\partial u_x}) - \frac{\partial F}{\partial u} \tag{80}$$

The energy functional will keep decreasing. For the 2D case, we have

$$\frac{\partial u}{\partial t} = \frac{d}{dx}(\frac{\partial F}{\partial u_x}) + \frac{d}{dy}(\frac{\partial F}{\partial u_y}) - \frac{\partial F}{\partial u} \tag{81}$$

Ex2: Then the gradient descent flow method to Ex1 is

$$\frac{\partial u}{\partial t} = div(\rho'(|\nabla u|)\frac{\nabla u}{|\nabla u|}) \tag{82}$$

Thus, intuitively, when taking $\rho(x) = x^2$. We obtain the gradient descent flow method is

$$\frac{\partial u}{\partial t} = div(\nabla u)$$

which is the linear diffusion.