PGMO Lecture: Vision, Learning and Optimization

1. Introduction

Thomas Pock

Institute of Computer Graphics and Vision

February 19, 2020

Outline of the lecture (1)

1. **Introduction:** Bayesian inference, image statistics, prior and likelihood modeling, total variation, ROF model, compressed sensing, sparse representations, LASSO model, connections, programming examples.
2. **Mathematical preliminaries:** Basic notions of convexity, subgradients, convex conjugate, infimal convolution, proximal map, duality, saddle-point problems.
3. **Proximal gradient methods:** implicit descent, proximal gradient, convergence rates, accelerations, line-search methods, mirror descent, programming examples.
4. **Primal-dual methods:** Convex-concave saddle point problems, primal-dual methods, convergence rates, line-search methods, accelerations, Augmented Lagrangian, ADMM, programming examples.
5. **Total Variation ++**: Advanced models based on the total variation, ℓ_1 fitting, entropy fitting, color-TV, total generalized variation, Potts model, functional lifting, convex relaxations, Mumford-Shah, curvature models, Total roto-translational, programming examples.

Outline of the lecture (2)

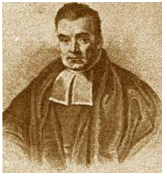
6. **Non-convex optimization:** proximal gradient methods in the non-convex setting, FISTA for non-convex problems, CoCaIn, iPALM, programming examples.
7. **Dynamic Programming:** Discrete labeling problems, relations to Markov random fields, dynamic programming on a chain, minimization vs. marginalization, semi-global matching, belief propagation, dual methods, dual minorize maximize, dynamic programming for continuous problems on a tree, accelerated dual block descent, programming examples.
8. **Learning better models**: Bilevel optimization, generalized TV model, Fields-of-experts model, early-stopping, optimal control, Total deep variation, learning with discrete labeling problems, programming examples.

Acknowledgments

- The main content of this lecture is based on joint work from the last 10+ years with great friends, colleagues, PostDocs and PhD students.
- **I'm particularly grateful to:** Antonin Chambolle, Karl Kunisch, Daniel Cremers, Christopher Zach, Kristian Bredies, Peter Ochs, Vladimir Kolmogorov, Shoham Sabach, Alexander Effland, Yura Malitsky, Alexander Shekhovtsov, Erich Kobler, Patrick Knöbelreiter, Kerstin Hammernik, Florian Knoll, Ricardo Otazo, Yunjin Chen, Gottfried Munda, Rene Ranftl, Manuel Werlberger, Markus Unger ...
- I also acknowledge funding from the Austrian Science Foundation (FWF) and the European Research Council (ERC).

Overview

Bayesian inference



- The very basis to statistical inference problems is formed by Bayes' theorem [Thomas Bayes 1701–1761]

$$p(u|d) = \frac{p(d|u)p(u)}{p(d)} \propto p(d|u)p(u)$$

- Treats both the measured data d as well as the solution u as random variables.
- Allows to asses the uncertainty of a certain solution u given the uncertainty of the measured data d.
- The posterior probability $p(u|d)$ of an unknown image u given data d is proportional to the likelihood $p(d|u)$ times the prior $p(u)$.
- The key challenge in the application of this formula to real-world problems is to find good mathematical models for $p(d|u)$ and $p(u)$.

Importance of the prior in Bayes theorem

- Humans have a certain (D)isease or are (H)ealthy with probability

$$p(D) = 20/10^5 = 0.0002, \quad p(H) = 1 - p(D) = 0.9998.$$

Importance of the prior in Bayes theorem

- Humans have a certain (D)isease or are (H)ealthy with probability

$$p(D) = 20/10^5 = 0.0002, \quad p(H) = 1 - p(D) = 0.9998.$$

- Assume, you have developed a (T)est which can detect the (D)isease with probability

$$p(T|D) = 0.95$$

Importance of the prior in Bayes theorem

- Humans have a certain (D)isease or are (H)ealthy with probability

$$p(D) = 20/10^5 = 0.0002, \quad p(H) = 1 - p(D) = 0.9998.$$

- Assume, you have developed a (T)est which can detect the (D)isease with probability

$$p(T|D) = 0.95$$

- However, the (T)est sometimes also fails and returns a positive result also for (H)ealthy people

$$p(T|H) = 0.01$$

Importance of the prior in Bayes theorem

- Humans have a certain (D)isease or are (H)ealthy with probability

$$p(D) = 20/10^5 = 0.0002, \quad p(H) = 1 - p(D) = 0.9998.$$

- Assume, you have developed a (T)est which can detect the (D)isease with probability

$$p(T|D) = 0.95$$

- However, the (T)est sometimes also fails and returns a positive result also for (H)ealthy people

$$p(T|H) = 0.01$$

- **What is the probability for an arbitrary human to have the (D)isease given a positive result of the (T)est?**

Importance of the prior in Bayes theorem

- Humans have a certain (D)isease or are (H)ealthy with probability

$$p(D) = 20/10^5 = 0.0002, \quad p(H) = 1 - p(D) = 0.9998.$$

- Assume, you have developed a (T)est which can detect the (D)isease with probability

$$p(T|D) = 0.95$$

- However, the (T)est sometimes also fails and returns a positive result also for (H)ealthy people

$$p(T|H) = 0.01$$

- **What is the probability for an arbitrary human to have the (D)isease given a positive result of the (T)est?**

- Using Bayes theorem, we can now evaluate the probability of a human to have the disease give a positive test:

$$p(D|T) = \frac{p(D)P(T|D)}{p(D)P(T|D) + p(H)P(T|H)} = \frac{0.0002 \times 0.95}{0.0002 \times 0.95 + 0.9998 \times 0.01} =$$

Importance of the prior in Bayes theorem

- Humans have a certain (D)isease or are (H)ealthy with probability

$$p(D) = 20/10^5 = 0.0002, \quad p(H) = 1 - p(D) = 0.9998.$$

- Assume, you have developed a (T)est which can detect the (D)isease with probability

$$p(T|D) = 0.95$$

- However, the (T)est sometimes also fails and returns a positive result also for (H)ealthy people

$$p(T|H) = 0.01$$

- **What is the probability for an arbitrary human to have the (D)isease given a positive result of the (T)est?**

- Using Bayes theorem, we can now evaluate the probability of a human to have the disease give a positive test:

$$p(D|T) = \frac{p(D)P(T|D)}{p(D)P(T|D) + p(H)P(T|H)} = \frac{0.0002 \times 0.95}{0.0002 \times 0.95 + 0.9998 \times 0.01} = 0.0186 \approx 2\%$$

Basic scientific questions

- The Bayesian formula naturally leads to the following questions:

- **Modeling:** How to model the a-priori distribution $p(u)$ and the likelihood distribution $p(d|u)$? Convex/non-convex models, learning, forward operator, noise, ...
- **Inference:** The Bayesian approach delivers a complete posterior distribution $p(u|d)$. Expectation, MAP estimation, ...

- Both questions are subject to intense current research and we are still far away from having good general answers.
- Moreover, both issues, modeling and inference usually have to be considered in a holistic way.

Exponential distributions

- Let us consider the following exponential distributions, well known from statistical mechanics [Gibbs, 1889]

$$p(u) \propto \exp\left(-\mathcal{R}(u)/T\right), \quad p(d|u) \propto \exp\left(-\mathcal{D}(d|u)/T\right),$$

  where T is the temperature (variance) of the configuration.
- $\mathcal{R}(u)$ is called prior, regularization, or smoothness term.
- $\mathcal{D}(d|u)$ is called data fidelity or data fitting term.

Exponential distributions

- Let us consider the following exponential distributions, well known from statistical mechanics [Gibbs, 1889]

$$p(u) \propto \exp\left(-\mathcal{R}(u)/T\right), \quad p(d|u) \propto \exp\left(-\mathcal{D}(d|u)/T\right),$$

  where T is the temperature (variance) of the configuration.
- $\mathcal{R}(u)$ is called prior, regularization, or smoothness term.
- $\mathcal{D}(d|u)$ is called data fidelity or data fitting term.

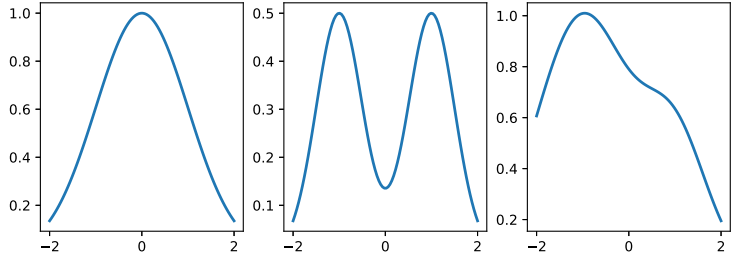
- Bayes' theorem can be re-written as

$$p(u|d) \propto \exp\left(-\mathcal{E}(u|d)/T\right),$$

  where $\mathcal{E}(u|d) = \mathcal{R}(u) + \mathcal{D}(d|u)$ is the "energy" of the configuration $(u|d)$.

Inference

- The Bayesian approach provides a complete posterior distribution $p(u|d)$, but which u should be selected?
- Consider the following three simple posterior distributions



- Unimodal, symmetric: Max $=$ Mean $=$ Median
- Bimodal, symmetric: Max $\neq$ Mean $=$ Median
- Unimodal, non-symmetric: Max $\neq$ Mean $\neq$ Median
- No clear answer possible ...

- General form: Bayesian estimator:

$$\hat{u} = \arg\min_u \int_V \ell(u, v) p(v|d) \, \mathrm{d}\mu(v)$$

Computing the expectation

- Select the sample u that minimizes the weighted squared loss function $\ell(u, v) = \|u - v\|^2$ to all possible images v

$$\bar{u} = \arg\min_u \int_V p(v|d) \, \|u - v\|^2 \, \mathrm{d}\mu(v)$$

- The solution is given by the weighted average (expectation)

$$\bar{u} = \frac{\int_V p(v|d) \, v \, \mathrm{d}\mu(v)}{\int_V p(v|d) \, \mathrm{d}\mu(v)} = \int_V p(v|d) \, v \, \mathrm{d}\mu(v)$$

- In practice hard to compute, because the integral over all possible images v cannot be evaluated.

Computing the maximum a-posterior estimate (MAP)

- Instead of a squared distance, choose a $0-1$ loss function $\ell(u,v) = \|u-v\|_0 = 0$ if $u = v$ and $\|u-v\|_0 = 1$ if $u \neq v$

$$u^* = \arg\min_u \int_V p(v|d)\, \|u-v\|_0 \; \mathrm{d}\mu(v)$$

- This loss function selects the sample u with the highest posterior probability and hence

$$u^* = \arg\max_u p(u|d) \overset{-\log(\cdot)}{=} \arg\min_u \mathcal{E}(u|d)$$

- Hence, the MAP estimate is equivalent to minimizing the energy $\mathcal{E}(u|d)$.

A statement of Euler from 1744

- *Cum enim mundi universi fabrica sit perfectissima atque a Creatore sapientissimo absoluta, nihil omnino in mundo contingit, in quo non maximi minimive ratio quaepiam eluceat; quamobrem dubium prorsus est nullum, quin omnes mundi effectus ex causis finalibus ope methodi maximorum et minimorum aeque feliciter determinari queant, atque ex ipsis causis efficientibus.*
- *Nothing in the world takes place without optimization, and there is no doubt that all aspects of the world that have a rational basis can be explained by variational methods.*[a]

[a] From the book: Optimization Stories

What is so difficult about images?

- Let us consider discrete images u of size $n \times n$ pixels, where each pixel can take one of 256 distinct gray values.

$n \times n$	1×1	2×2	4×4	8×8	16×16
$\#u$	$2.6 \cdot 10^2$	$4.3 \cdot 10^9$	$3.4 \cdot 10^{38}$	$1.3 \cdot 10^{154}$	$3.2 \cdot 10^{616}$

What is so difficult about images?

- Let us consider discrete images u of size $n \times n$ pixels, where each pixel can take one of 256 distinct gray values.

$n \times n$	1×1	2×2	4×4	8×8	16×16
$\#u$	$2.6 \cdot 10^2$	$4.3 \cdot 10^9$	$3.4 \cdot 10^{38}$	$1.3 \cdot 10^{154}$	$3.2 \cdot 10^{616}$

- The number of atoms in the observable universe is estimated to be between 10^{78} and 10^{82}.

What is so difficult about images?

- Let us consider discrete images u of size $n \times n$ pixels, where each pixel can take one of 256 distinct gray values.

$n \times n$	1×1	2×2	4×4	8×8	16×16
$\#u$	$2.6 \cdot 10^2$	$4.3 \cdot 10^9$	$3.4 \cdot 10^{38}$	$1.3 \cdot 10^{154}$	$3.2 \cdot 10^{616}$

- The number of atoms in the observable universe is estimated to be between 10^{78} and 10^{82}.
- The game-tree complexity of chess (Shannon number) is estimated to be at least 10^{123}.

What is so difficult about images?

- Let us consider discrete images u of size $n \times n$ pixels, where each pixel can take one of 256 distinct gray values.

$n \times n$	1×1	2×2	4×4	8×8	16×16
$\#u$	$2.6 \cdot 10^2$	$4.3 \cdot 10^9$	$3.4 \cdot 10^{38}$	$1.3 \cdot 10^{154}$	$3.2 \cdot 10^{616}$

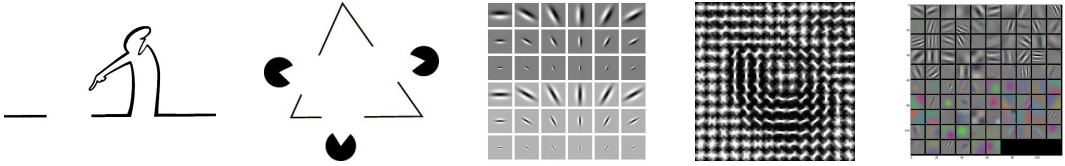
- The number of atoms in the observable universe is estimated to be between 10^{78} and 10^{82}.
- The game-tree complexity of chess (Shannon number) is estimated to be at least 10^{123}.

- Only a small fraction of all possible images represent natural images.
- How does the distribution of images look like?

Overview

Edges

- **Edges** are among the most important features in images
- Image understanding relies on abstract discontinuity information
- Most successful image descriptors are based on intensity gradients
- First layers in deep convolutional networks represent edge detectors



(a) (b) (c) (d) (e)

- In (b), meaningful edges are absent yet "visible"... (would deep learning solve this??)
- What about the statistics of small image patches?

Image gradients

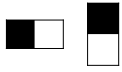
- Treating images $u : \Omega \to \mathbb{R}$ as continuous (differentiable) functions, edges correspond to strong image gradients:

$$\nabla u = \begin{pmatrix} \partial_{x_1} u \\ \partial_{x_2} u \end{pmatrix}, \quad |\nabla u| = \sqrt{(\partial_{x_1} u)^2 + (\partial_{x_2} u)^2}.$$

- Treating images as discrete arrays $u \in \mathbb{R}^{m \times n}$, image gradients are approximated using finite differences, e.g.

$$\nabla u \approx \begin{pmatrix} u_{i+1,j} - u_{i,j} \\ u_{i,j+1} - u_{i,j} \end{pmatrix}, \quad |\nabla u| \approx \sqrt{(u_{i+1,j} - u_{i,j})^2 + (u_{i,j+1} - u_{i,j})^2}$$

- Observe that the finite differences approximations can be computed by convolving the image with small filters kernels

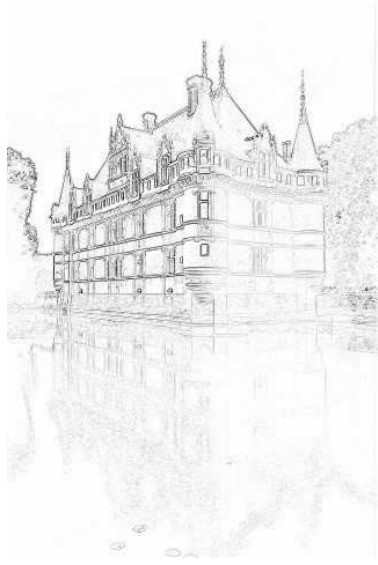


black=-1, white=+1

- What is the statistics of edges in natural images?

Example

Image u

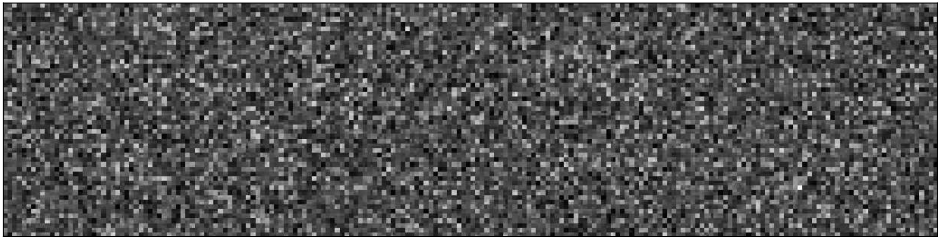


Edges $|\nabla u|$

Overview

Statistics of natural images

- As a first step, let us investigate the statistics of 2×2 patches $x_i \in \mathbb{R}^{2\times 2}$, $i = 1, ..., N$, extracted from natural images.



- We project each patch to the basis functions of a 2×2 DCT transform:

f_1 f_2 f_3 f_4

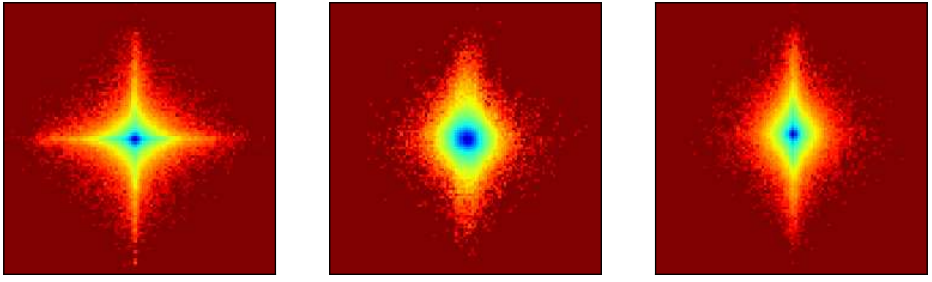
- f_1 captures the average gray value and hence can be ignored.
- $f_2, ..., f_4$ act as "edge detectors" and hence carry the statistics of edges.

Statistics of natural images

- The projected patches are computed by

$$y_i = \left(\langle x_i,\, f_2\rangle\,, \langle x_i,\, f_3\rangle\,, \langle x_i,\, f_4\rangle\right) \in \mathbb{R}^3.$$

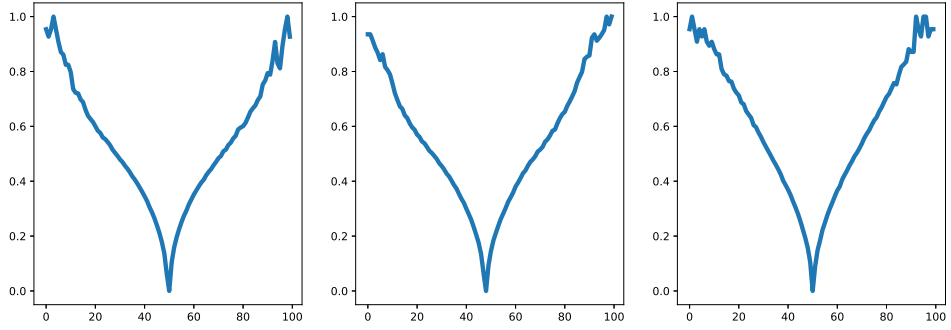
- The resulting negative log density function $-\log p(y)$ of all projected patches is very non-Gaussian:



Slices of the 3D negative log density

Marginals

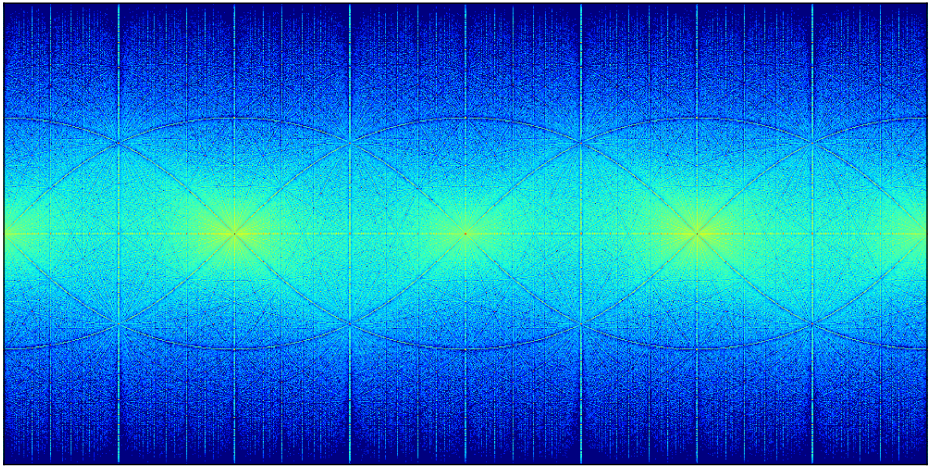
- Marginal distributions of the edge filters $f_2, ..., f_4$



- Observe that the marginal distributions are heavy-tailed, which means that most patches are in non-edge regions.

Contrast normalization

- Performing a contrast normalization $\tilde{y}_i = y_i/|y_i|_2$ projects the 3D density onto the surface of a sphere.



Unrolled surface

- Has a surprisingly regular and fine structure.

Modeling the prior term

- Assume that x is an image patch and we have a set of filters $F = (f_1, ..., f_K)$ such that $Fx = (\langle x, f_1 \rangle, ..., \langle x, f_K \rangle)$.
- As discussed before, a widely used approach is to assume the following factorized exponential distribution:

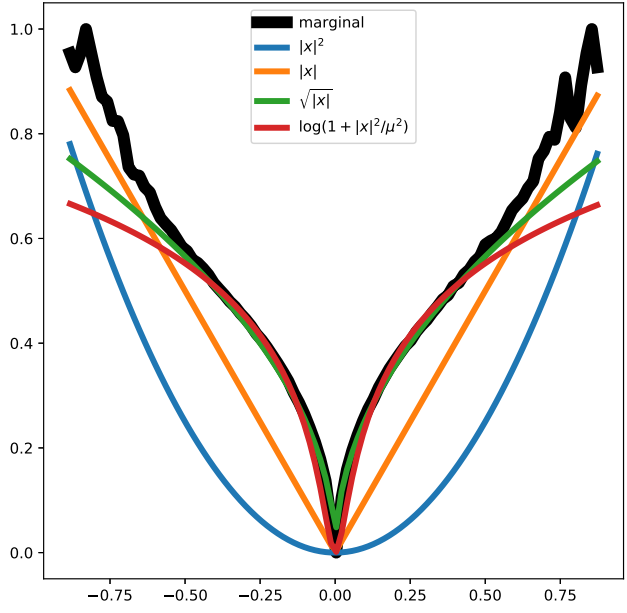
$$p(x) \propto \exp(-\rho(Fx)) \Longleftrightarrow R(x) = \rho(Fx)$$

  where ρ is a potential function.
- The potential function can work on the whole vector Fx or one can also assume that it factorizes over its marginals

$$p(x) \propto \prod_{k=1}^{K} p_k(x) = \prod_{k=1}^{K} \exp(-\rho_k(\langle x, f_k \rangle)) \Longleftrightarrow R(x) = \sum_{k=1}^{K} \rho_k(\langle x, f_k \rangle)$$

Potential functions for the marginals

- The potential functions should match the negative log probability of the true statistics.



Convex or non-convex?

- The quadratic function $\rho(t) = |t|^2$ (corresponding to a Gaussian) is completely off.
- The absolute function $\rho(t) = |t|$ is much better but still off.
- The best matches are obtained for $\rho(t) = \sqrt{|t|}$ or $\rho(t) = \log(1 + |t|^2/\mu^2)$, but the functions are non-convex.

- A good trade-off between model fit and complexity seems to be $\rho(t) = |t|$.
- Besides 1D potential functions, potential functions acting on several features can also be considered, for example the 2-norm $\rho(t) = |t|_2 = \sqrt{t_1^2 + ..., t_n^2}$.

- Later we will see that we can obtain significantly better results for non-convex (learned) potential functions.

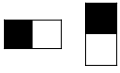
From patch priors to image priors

- So far we have restricted our statistical model to small patches $x \in \mathbb{R}^{2\times 2}$.
- How can we apply this model to whole images $u \in \mathbb{R}^{m\times n}$?
- Since images are usually assumed to be translational invariant, we one can represent the filter correlations $\langle c, f_k \rangle$ by means of 2D convolutions $f_k * u$.

- Adopting the convex potential functions $\rho(t) = |t|$ leads to the following fundamental image prior model:

$$R(u) = \sum_{k=1}^{K} \|f_k * u\|_1 = \sum_{k=1}^{K} \sum_{i,j} |(f_k * u)_{i,j}|.$$

- A very fundamental image prior is obtained if f_1 and f_2 are discrete derivative filters



Overview

Modeling the data term

- The structure of the data term is basically derived from a linear (or non-linear) forward operator A and the noise statistics.
- Assume the following linear image formation process

$$d = Au + n,$$

where d is the given data, A is a linear operator and $n \sim \mathcal{N}(0, \sigma^2)$ is assumed to be i.i.d. zero-mean Gaussian noise with variance σ^2, then the data term is chosen as the negative log probability of a Gaussian distribution:

$$\mathcal{D}(d|u) = \frac{1}{2} \|Au - d\|^2,$$

where A is the linear forward operator of the problem.

- In case of other noise statistics, other data fidelity terms can be used, for example the ℓ_1 norm (Laplacian noise)

$$\mathcal{D}(d|u) = \|Au - d\|_1,$$

or the entropy (multiplicative Gamma noise)

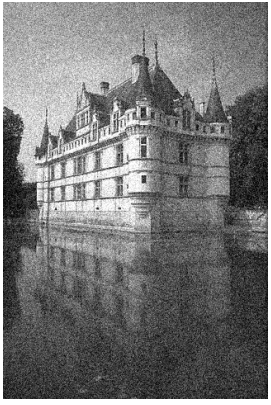
$$\mathcal{D}(d|u) = \sum_p (Au)_p - d_p \log((Au)_p), \quad (Au)_p > 0.$$

Example: Additive Gaussian noise

- In case of additive Gaussian noise, the given data d is the noisy image and the linear operator $A = \mathrm{id}$.
- The data fidelity term is then given by

$$D(d|u) = \frac{1}{2} \|u - d\|^2$$

Clean image



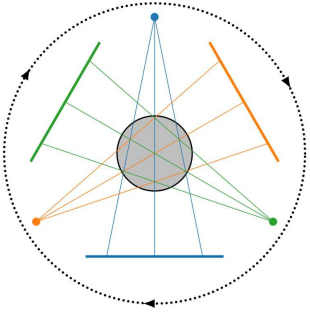
Noisy image

Computed tomography (CT)

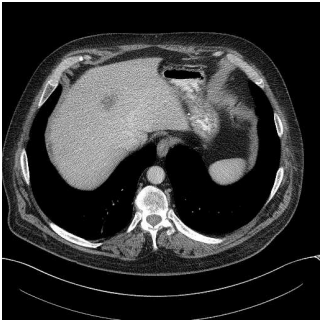
- In case of CT, the linear operator is given by the Radon (Ray) transform

$$d_\ell \approx (Au)(\ell) = \int_\ell u \, \mathrm{d}s.$$

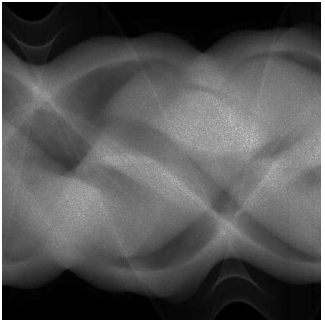
- In practice only a finite number of lines ℓ can be measured and the measurements are degraded by noise.



Working procedure of CT



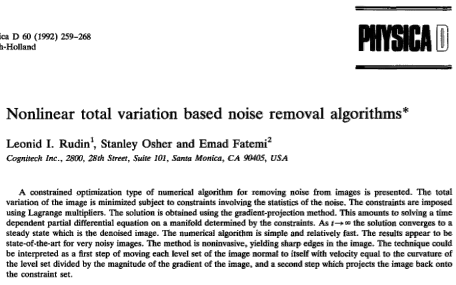
CT image
Source: Jonas Adler



Line measurements (sinogram)

Overview

The ROF model



Physica D 60 (1992) 259–268
North-Holland

PHYSICA D

Nonlinear total variation based noise removal algorithms*

Leonid I. Rudin[1], Stanley Osher and Emad Fatemi[2]
Cognitech Inc., 2800, 28th Street, Suite 101, Santa Monica, CA 90405, USA

A constrained optimization type of numerical algorithm for removing noise from images is presented. The total variation of the image is minimized subject to constraints involving the statistics of the noise. The constraints are imposed using Lagrange multipliers. The solution is obtained using the gradient-projection method. This amounts to solving a time dependent partial differential equation on a manifold determined by the constraints. As $t \to \infty$ the solution converges to a steady state which is the denoised image. The numerical algorithm is simple and relatively fast. The results appear to be state-of-the-art for very noisy images. The method is noninvasive, yielding sharp edges in the image. The technique could be interpreted as a first step of moving each level set of the image normal to itself with velocity equal to the curvature of the level set divided by the magnitude of the gradient of the image, and a second step which projects the image back onto the constraint set.

- The ROF model is defined as the following minimization problem

$$\min_u \lambda \int_\Omega |Du| + \frac{1}{2} \|u - f\|^2, \ \lambda > 0$$

- Defines "the" prototypical optimization problem in mathematical imaging
- It consists of a total variation regularizer and a quadratic data fidelity term
- Gives a good trade-off between simplicity of the model and denoising quality
- Before we can solve it, we need to consider a discrete version of the total variation

The discrete total variation

- We consider a scalar-valued digital image $u \in \mathbb{R}^{m \times n}$ of $m \times n$ pixels
- We define a finite differences operator $\mathrm{D} : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n \times 2}$

$$(\mathrm{D}u)_{i,j,1} = \begin{cases} u_{i+1,j} - u_{i,j} & \text{if } 1 \leq i < m, \\ 0 & \text{else,} \end{cases}$$

$$(\mathrm{D}u)_{i,j,2} = \begin{cases} u_{i,j+1} - u_{i,j} & \text{if } 1 \leq j < n, \\ 0 & \text{else.} \end{cases}$$

- We will later also need the operator norm $\|\mathrm{D}\|$ which is estimated as $\|\mathrm{D}\| \leq \sqrt{8}$.
- The discrete total variation is defined as

$$\|\mathrm{D}u\|_{p,1} = \sum_{i=1,j=1}^{m,n} |(\mathrm{D}u)_{i,j}|_p = \sum_{i=1,j=1}^{m,n} \left((\mathrm{D}u)_{i,j,1}^p + (\mathrm{D}u)_{i,j,2}^p \right)^{1/p},$$

  that is, the ℓ_1-norm of the p-norm of the pixelwise image gradients.
- For $p = 1$ we obtain the anisotropic total variation and if $p = 2$ we obtain the isotropic total variation.

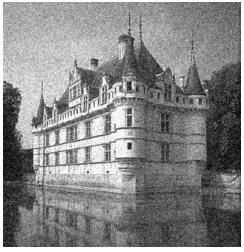
The discrete ROF model

- The discrete ROF model is defined as

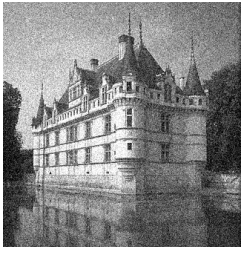
$$\min_u \lambda \|\mathrm{D}u\|_{p,1} + \frac{1}{2}\|u - d\|_2^2,$$

  where $d \in \mathbb{R}^{m \times n}$ is the given noisy image and $\lambda > 0$ is a trade-off parameter.
- It is one of the most fundamental models in mathematical image processing.
- Has been extensively studied and extended in various ways (more on that later).
- Its main advantage is that that it can identify and preserve the most fundamental discontinuities (edges) in images.

Total variation minimization



u, $\lambda = 0.01$

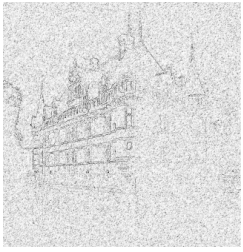


d

$(u_{i+1,j} - u_{i,j})$

$(u_{i,j+1} - u_{i,j})$

$$\min_u \lambda \|\mathrm{D}u\|_{2,1} + \frac{1}{2}\|u - d\|_2^2 .$$

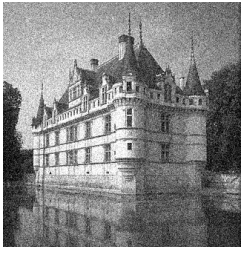


$\sqrt{(u_{i+1,j} - u_{i,j})^2 + (u_{i,j+1} - u_{i,j})^2}$

$\overset{\Sigma_{i,j}}{\rightsquigarrow} TV \approx 23501.84$

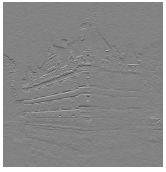
Total variation minimization

u, $\lambda = 0.05$



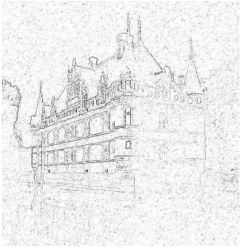
d

$(u_{i+1,j} - u_{i,j})$



$(u_{i,j+1} - u_{i,j})$

$$\min_u \lambda \|\mathrm{D}u\|_{2,1} + \frac{1}{2}\|u - d\|_2^2 .$$

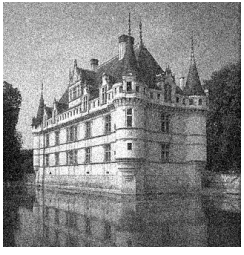


$\sqrt{(u_{i+1,j} - u_{i,j})^2 + (u_{i,j+1} - u_{i,j})^2}$

$\overset{\sum_{i,j}}{\rightsquigarrow} TV \approx 7916.47$

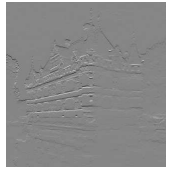
Total variation minimization

u, $\lambda = 0.1$



d

$(u_{i+1,j} - u_{i,j})$



$(u_{i,j+1} - u_{i,j})$

$$\min_u \lambda \|\mathrm{D}u\|_{2,1} + \frac{1}{2}\|u - d\|_2^2 .$$

$\sqrt{(u_{i+1,j} - u_{i,j})^2 + (u_{i,j+1} - u_{i,j})^2}$

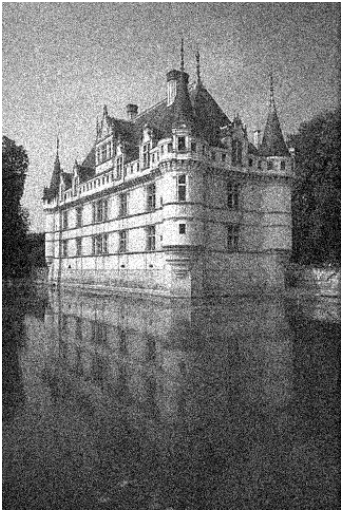
$\overset{\sum_{i,j}}{\rightsquigarrow} TV \approx 2987.83$

Some properties of the discrete total variation

- From a sparsity point of view, the total variation induces sparsity in the gradients of the image, hence, it favors piecewise constant images
- This property is known as staircasing effect, which is often considered as a drawback for certain applications
- The case $p = 1$ allows for quite effective splitting techniques but favors edges to be aligned with the grid
- The case $p = 2$ can also be considered as a simple form of group sparsity, grouping together the spatial derivatives in each dimension
- The isotropic variant does not exhibit a grid bias and hence is often preferred in practice

Example

(a) Original image



(b) Noisy image d

(c) Denoised image u

The TV-ℓ_1 model

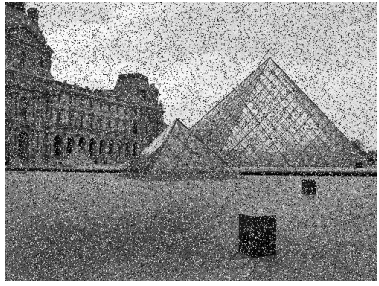
- The ROF model performs well in case of Gaussian noise, but it performs very bad in case of impulsive noise / outliers
- The TV-ℓ_1 model is obtained by replacing the quadratic data fitting term by an ℓ_1 norm data fitting term

$$\min_u \lambda \int_\Omega |Du| + \|u - f\|_1$$

- Has a lot of nice properties:
- The model is contrast invariant: If u^* is a minimizer for f then cu^* is also a minimizer for cf, with $c > 0$
- The model has a morphological property, i.e. it can be used for scale selection

Example

(a) Original image



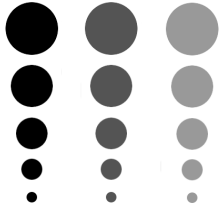
(b) Noisy image

(c) TV-ℓ_1

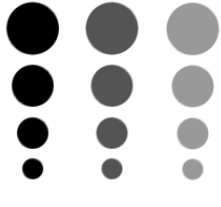
(d) ROF

Scale selection

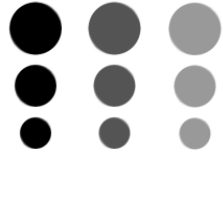
TV-ℓ_1



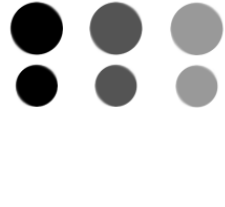
(a) $\lambda = 0$



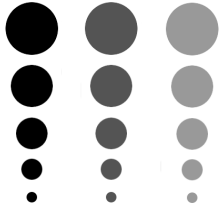
(b) $\lambda = 4$



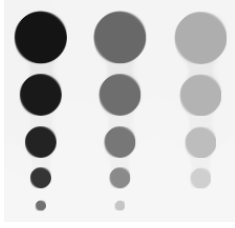
(c) $\lambda = 5$



(d) $\lambda = 8$

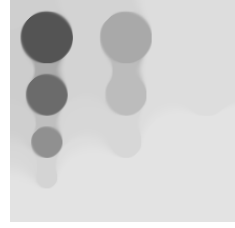


(e) $\lambda = 0$



(f) $\lambda = 1$

(g) $\lambda = 2$



(h) $\lambda = 4$

ROF

Multiplicative noise

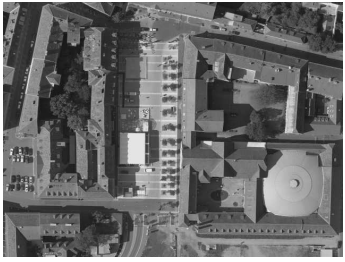
- In some image acquisition techniques the type of noise is not additive but multiplicative. The image degradation process can be written as

$$f(x) = u(x) \cdot n(x),$$

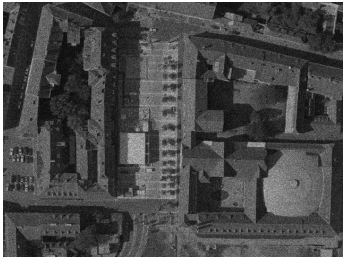
  where n is the noise component which has mean 1 which is usually modeled as a Gamma- or Poisson distribution.
- Examples include ultrasound (US) images, single particle emission computed tomography (SPECT), Synthetic Aperture Radar (SAR) images, etc.
- In such cases, it turns out that an TV regularized model with an entropy-type data term performs much better

$$\min_{u>0} \lambda \int_\Omega |Du| + \int_\Omega u(x) - f(x) \log u(x) \, \mathrm{d}x$$

Example: Poisson noise

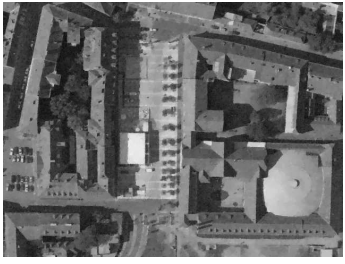


(a) original image



(b) noisy image

(c) ROF, $\lambda = 10$



(d) TV-entropy, $\lambda = 7$

Deblurring

- The ROF model can be modified such that the model can be used for image deblurring with a known point spread function

$$\min_u \int_\Omega \lambda |Du| + \frac{1}{2} \|a * u - f\|^2$$

- The point spread function a is usually a small filter kernel that has been derived from the imaging system

Gaussian

Motion

Example

(a) Original image

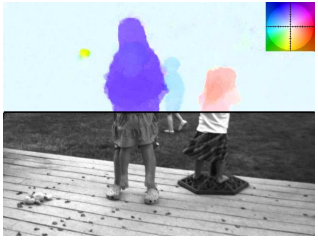
(b) Blurry and noisy image

(c) Deblurred, no regularization

(d) Deblurred, using TV

Further applications

- Total variation based models can be used to solve even more inverse problems in imaging:



(a) Motion

(b) Stereo

(c) Segmentation

`tv.ipynb`

Overview

Shannon-Nyquist sampling theorem

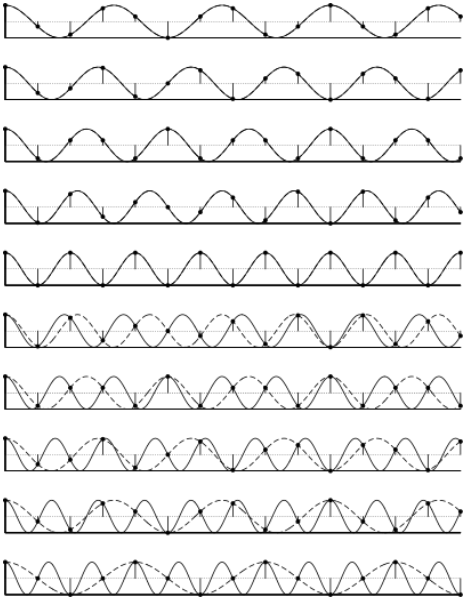
- In the field of digital signal processing, the sampling theorem is a fundamental bridge between continuous-time signals and discrete-time signals
- It establishes a sufficient condition for a sample rate that avoids aliasing

$$f_s > 2f_{\max},$$

  where f_s is the sampling frequency and $f_{\max}$ is the maximal frequency of the signal to be sampled.
- In practice usually a slightly larger factor (e.g. 2.2) has to be chosen (due to the inexactness of low-pass filters).

Example

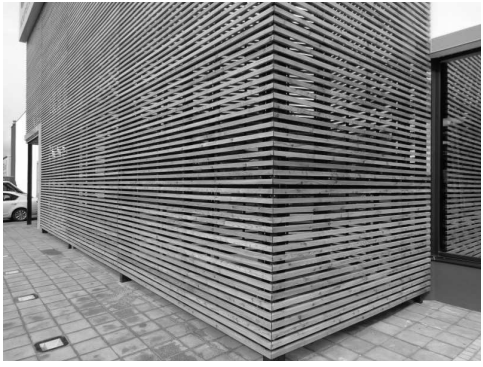
- The sampling points are take always at the same positions
- The frequency increases from top to bottom
- After violating the Shannon-Nyquist sampling theorem, aliasing occurs
- The dashed curves are possible signals with the same sampling points

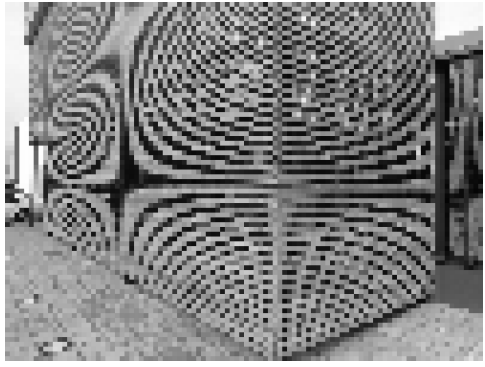


Source: Wikipedia

Example

Example: Aliasing in $8\times$ undersampled image





Compressed sensing

- Compressed sensing (CS) is a signal processing technique for efficiently acquiring and reconstructing a signal
- It is based on finding solutions to underdetermined linear systems
- The underlying principle is that the sparsity of a signal can be exploited to recover it from far fewer samples than required by the Shannon-Nyquist sampling theorem.
- Sparsity can be understood in different ways but always refers to some regularity assumption on the signal.
- Sparsity in a basis: Only a few basis atoms of a certain basis (e.g. dictionary) have to be used to reconstruct the signal
- Sparsity in a transform: After applying a transform to the signal (e.g. the gradient operator) the signal becomes sparse (e.g. the total variation).

Solutions of underdeterminded systems of equations

- At the very heart of compressed sensing is computing solutions of underdetermined systems of equations.
- Let us consider the following underdetermined system of equations of the form

$$Ax = b$$

- b is a $m \times 1$ measurement vector
- x is the $n \times 1$ unknown signal
- A is the $m \times n$ system matrix, with $m < n$
- How can we solve the underdetermined system of equations?

Regularization

- Let us consider the regularized problem

$$\min_x J(x) \quad \text{subject to} \quad Ax = b$$

- A first simple choice is the squared ℓ_2 distance $J(x) = \|x\|_2^2$
- The unique solution $\hat{x}$ of the problem is then given by

$$\hat{x}_2 = A^T(AA^T)^{-1}b,$$

  which is exactly the pseudo-inverse of A.
- The quadratic regularization tries to find a solution $\hat{x}$ that has the smallest ℓ_2 norm.

Sparsity

- Another form of regularization that received a lot of attention during the last years is based on sparsity
- The idea is that the underlying "dimension" of a signals' complexity is small if represented in a suitable basis
- A simple and intuitive form of sparsity is given by the ℓ_0 (pseudo) norm of a vector x

$$J(x) = \|x\|_0 = \#\{i : x_i \neq 0\},$$

and hence $\|x\|_0 < n$ if x is sparse.

- Hence we consider the following problem

$$\min_x \|x\|_0 \quad \text{subject to} \quad Ax = b,$$

which is known as "Basis Pursuit" [Chen, Donoho, '94].

Convex relaxation

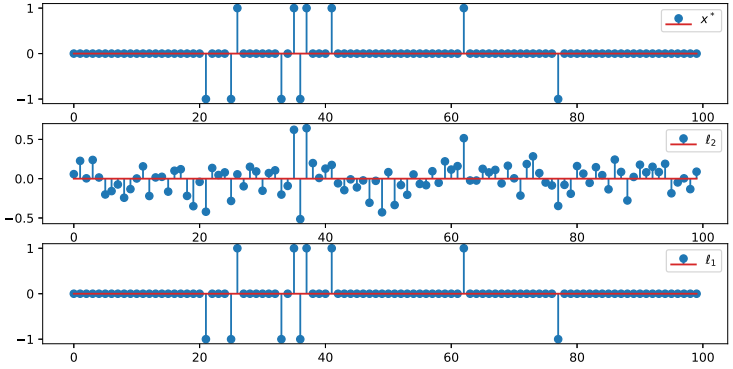
- The previous problem is NP-hard and hence very hard to solve if the degree of sparsity is not very small
- A simple idea is to replace the ℓ_0 pseudo norm by its closest convex approximation, i.e. $J(x) = \|x\|_1$:

$$\min_x J(x) = \|x\|_1 \quad \text{subject to} \quad Ax = b,$$

- This problem can actually be solved using convex optimization algorithms
- Under certain circumstances, the solution of the convex ℓ_1 problem yields the same sparse solution as the solution of the ℓ_0 problem

ℓ_1 versus ℓ_2

- Comparison between ℓ_2 regularization and ℓ_1 regularization to recover a sparse signal (sparsity $= 0.1$).
- The signal is of length $n = 100$, the number of measurements (random projections) is $m = 30$.
- The ℓ_1 minimization approach can successfully recover the signal.



`sparsity.ipynb`

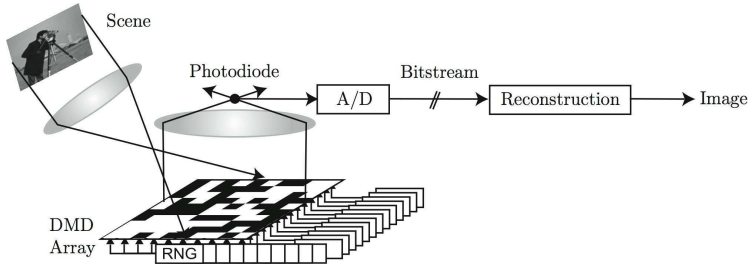
Sparsity in a transform

- In case the signal x is not sparse but it is expected to be sparse in some transform, it makes sense to choose a regularizer $J(x) = \|Lx\|_1$:

$$\min_x \|Lx\|_1 \quad \text{subject to} \quad Ax = b,$$

  where L is some "sparsifying" linear transform.
- In imaging, the total variation has been turned out to be a very successful sparsifying transform, which is obtained by setting $L = \mathrm{D}$.

Example: One-Pixel camera (Rice University):



Source: http://elec424.rice.edu/spc/index.html

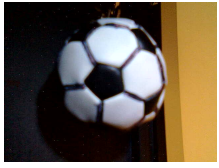
- The digital micro mirror device (DMD) generates a number of "projections" $a_i^T x$ of the image that are measured by the single photodiode.

$$a_i^T x = b_i, \quad i = 1, ..., m \Longleftrightarrow Ax = b$$

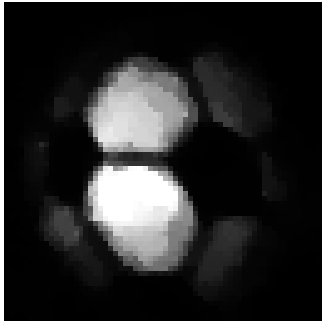
- The task of compressed sensing is now to reconstruct the image from a number of such single-pixel measurements b_i by solving

$$\min_x \|\mathrm{D}x\|_1 \quad \text{subject to} \quad Ax = b,$$

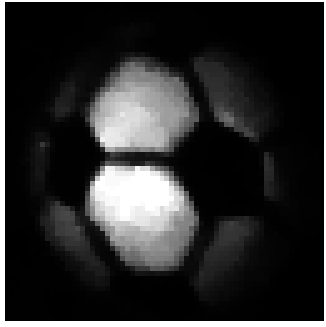
Results



Original object



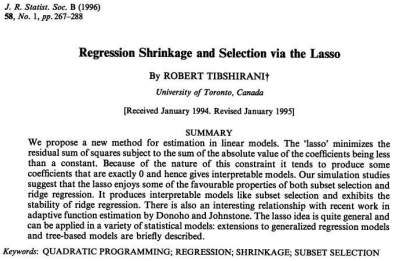
800 meas. (20%)



1600 meas. (40%)

Source: http://elec424.rice.edu/spc/index.html

The LASSO model



J. R. Statist. Soc. B (1996)
58, *No.* 1, *pp.* 267–288

Regression Shrinkage and Selection via the Lasso

By ROBERT TIBSHIRANI†
University of Toronto, Canada

[Received January 1994. Revised January 1995]

SUMMARY
We propose a new method for estimation in linear models. The 'lasso' minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficients that are exactly 0 and hence gives interpretable models. Our simulation studies suggest that the lasso enjoys some of the favourable properties of both subset selection and ridge regression. It produces interpretable models like subset selection and exhibits the stability of ridge regression. There is also an interesting relationship with recent work in adaptive function estimation by Donoho and Johnstone. The lasso idea is quite general and can be applied in a variety of statistical models: extensions to generalized regression models and tree-based models are briefly described.

Keywords: QUADRATIC PROGRAMMING; REGRESSION; SHRINKAGE; SUBSET SELECTION

- In case there is noise in the measurement, we can replace the equality in the constraint by an inequality constraint, leading to

$$\min_x \|x\|_1 \quad \text{subject to} \quad \|Ax - b\|^2 \leq \sigma^2,$$

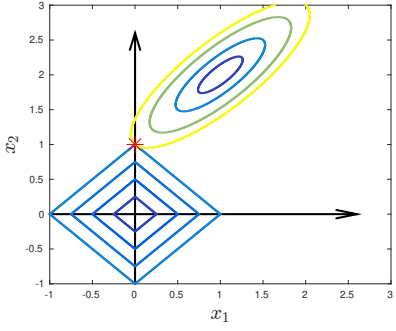
  where $\sigma > 0$ is an estimate of the noise level.
- This problem can equivalently be written as the unconstrained optimization problem

$$\min_x \|x\|_1 + \frac{\lambda}{2} \|Ax - b\|^2,$$

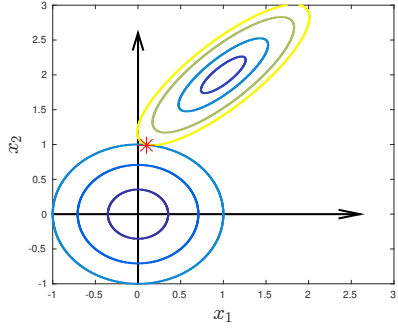
  where $\lambda > 0$ is a suitable Lagrange multiplier.
- This model is known as the "Lasso" (Least absolute shrinkage and selection operator)

Basic functionality

- In statistics, the Lasso model is used to perform linear regression and regularization, order to improve the prediction accuracy of a statistical model
- Sparsity in the Lasso model has a nice geometric interpretation why the ℓ_1 norm leads to sparse solutions



$f(x) = \|\cdot\|_1$



$f(x) = \|\cdot\|_2^2$

Other sparsity inducing functions

Besides the ℓ_1 norm, there are other interesting sparsity inducing functions. Assume $x \in \mathbb{R}^{m \times n}$

- Mixed $\ell_{1,2}$ norm: $\|x\|_{2,1} = \sum_{i=1}^{m} |\sqrt{\sum_{j=1}^{n} |x_{i,j}|^2}|$ can be used to induce sparsity in groups of variables
- The nuclear norm $\|x\|_* = \sum_{i=1}^{\min\{m,n\}} \sigma_i(x)$ can be used to induce sparsity in the singular values of x which in turn imposes a low rank prior on x

Relations between Lasso and ROF

- Let us again consider the Lasso model with A being invertible

$$\min_x \|x\|_1 + \frac{\lambda}{2} \|Ax - b\|^2,$$

and let us perform the following change of variables:

$$y = Ax \Rightarrow x = A^{-1}y$$

- It follows we can write a new model in the y variable and letting $D = A^{-1}$:

$$\min_y \|Dy\|_1 + \frac{\lambda}{2} \|y - b\|^2,$$

- Leads to a regularizer that enforces sparsity in the transform D.
- This model has exactly the same for as the ROF model for total variation based image denoising.
- Usually A is not invertible (over- or underdetermined) and hence the relation does not hold exactly.

Synthesis vs. analysis

- Again consider the difference between the Lasso model (left) and the ROF model (right):

$$\min_x \|x\|_1 + \frac{\lambda}{2}\|Ax - b\|^2, \quad \min_y \|Dy\|_1 + \frac{\lambda}{2}\|y - b\|^2,$$

- In the Lasso model, the linear operator A "synthesizes" the signal while in the ROF model, the linear operator D "analyzes" the signal
- In general it is believed that the analysis-based model is richer.
- We can of course have both operators:

$$\min_x \|Dx\|_1 + \frac{\lambda}{2}\|Ax - b\|^2$$

- Used in all TV-regularized inverse problems (MRI reconstruction, deblurring, ...)

Example: Image compression using Lasso

- Let us consider the following Lasso problem

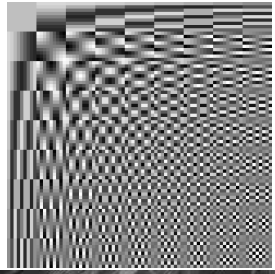
$$\min_x \lambda \|x\|_1 + \frac{1}{2} \|Ax - b\|^2,$$

- $b \in \mathbb{R}^{mm}$ is a (vectorized) image patches of size $m \times m$.
- $A \in \mathbb{R}^{mm \times mm}$ defines the basis which is given by the (vectorized) basis patches of size $m \times m$ of a 2D DCT
- $x \in \mathbb{R}^{mm}$ is the compressed patches in the basis A
- Assuming that $A^{-1} = A^*$ the problem can be solved in closed form using "soft-shrinkage":

$$x = \max(0, |A^* b| - \lambda) \cdot \operatorname{sgn}(A^* b)$$

`lasso.ipynb`

Results



sparsity= 0.13

sparsity= 0.05

Image denoising using wavelets

- We can also denoise a complete image using sparsity in the wavelet transform
- Consider again the Lasso problem

$$\min_d \lambda \|d\|_1 + \frac{1}{2} \left\| W^T(d, c) - f \right\|^2,$$

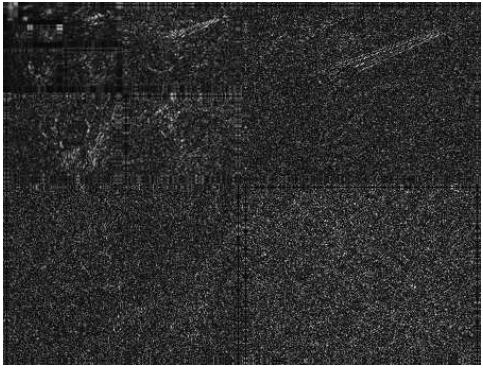
  where W defines the wavelet transform, f is the noisy image, d are the wavelet detail coefficients, and c is the approximation coefficient.
- Leads to an image which has a sparse wavelet coefficients
- Sparsity is not used in the approximation coefficients!

Example

- Clean and noisy image using $\sigma = 0.1$

Example

- Approximation and detail coefficients using biorthogonal wavelets before and after shrinkage



Before shrinkage

After shrinkage

Example

- Noisy image and denoising result ($\lambda = 0.1$)