Convex Optimization Theory and Applications

Topic 8 - Norm Approximation and Regularization

Li Li

Department of Automation,
Tsinghua University,

Fall, 2009-2021.

8.0. Outline

8.1. Norm Approximation

- 8.1.1 Linear Norm Approximation
- 8.1.2 Nonlinear Norm Approximation

8.2. l_1 Magic

- 8.2.1 l_0 Problem and Sparsity
- 8.2.2 l_1 Magic
- 8.2.3 Compressive Sensing

8.3. Regularization

- 8.3.1 Tikhonov Regularization
- 8.3.2 Regularization in Learning
- 8.3.3 Imaging the Invisiable

8.1. Norm Approximation

8.1.1 Linear Norm Approximation

$$\text{minimize} \quad \|Ax - b\| \tag{8.1}$$

where $A \in R^{m\times n}$ with $m \geq n$, $\|\cdot\|$ is a norm on R^m

interpretations of solution $x^* = \arg\min_x \|Ax - b\|$:
geometric: Ax^* is point in $R(A)$ closest to b
estimation: linear measurement model

$$y = Ax + v \tag{8.2}$$

y are measurements, x is unknown, v is measurement error

8.1. Norm Approximation

8.1.1 Linear Norm Approximation

Assumptions made in Galileo Galilei, *Dialogo sopra i due massimi systemi del mondo, tolemaico e copernicano*

1. Errors DO exist

2. Errors distributions are symmetric

3. The occurrence probability of a large error is smaller than the occurrence probability of a small error

8.1. Norm Approximation

8.1.1 Linear Norm Approximation

Least-square approximation ($\|\cdot\|_2$): solution satisfies normal equations

$$A^T A x = A^T b \tag{8.3}$$

where $x^* = (A^T A)^{-1} A^T b$ if $Rank(A) = n$

8.1. Norm Approximation

8.1.1 Linear Norm Approximation

Least-square approximation ($\|\cdot\|_2$): solution satisfies normal equations

$$A^T Ax = A^T b \qquad (8.3)$$

where $x^* = (A^T A)^{-1} A^T b$ if $Rank(A) = n$

Brief Proof: Because $f(x) = x^T A^T Ax - 2b^T Ax + b^T b$, a point x minimizes $f(x)$ if and only if

$$\nabla f(x) = 2A^T Ax - 2Ab^T = 0$$

8.1. Norm Approximation

8.1.1 Linear Norm Approximation

Chebyshev approximation ($\|\cdot\|_\infty$): can be solved as an LP

$$\text{minimize} \quad t \tag{8.4}$$

subject to $-t\cdot 1 \le Ax-b \le t\cdot 1$

Sum of absolute residuals approximation ($\|\cdot\|_1$): can be solved as an LP

$$\text{minimize} \quad 1^T y \tag{8.5}$$

subject to $-y \le Ax-b \le y$.

8.1. Norm Approximation

8.1.2 Nonlinear Norm Approximation

$$\text{minimize} \quad \phi(r_1)+\cdots+\phi(r_m) \qquad (8.6)$$

subject to $r = Ax - b$, where $A \in R^{m\times n}$, $\phi : R \to R$ is a convex penalty function

quadratic: $\phi(u) = u^2$

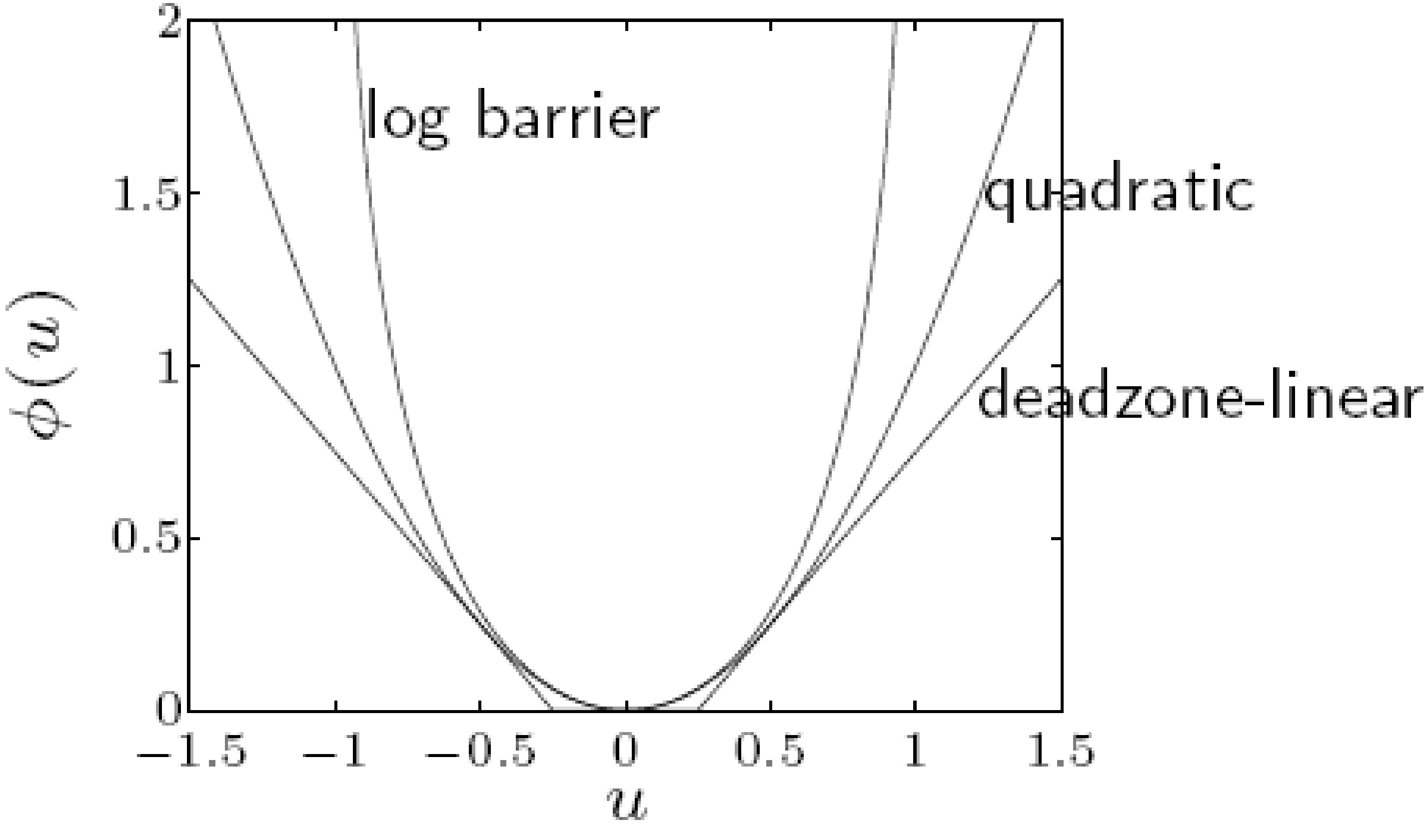
deadzone-linear with width a: $\phi(u) = \max\{0, |u| - a\}$

Log-barrier with limit a:

$$\phi(u) = \begin{cases} -a^2 \log(1-(u/a)^2) & |u| < a \\ \infty & otherwise \end{cases}$$

8.1. Norm Approximation

8.1.2 Nonlinear Norm Approximation

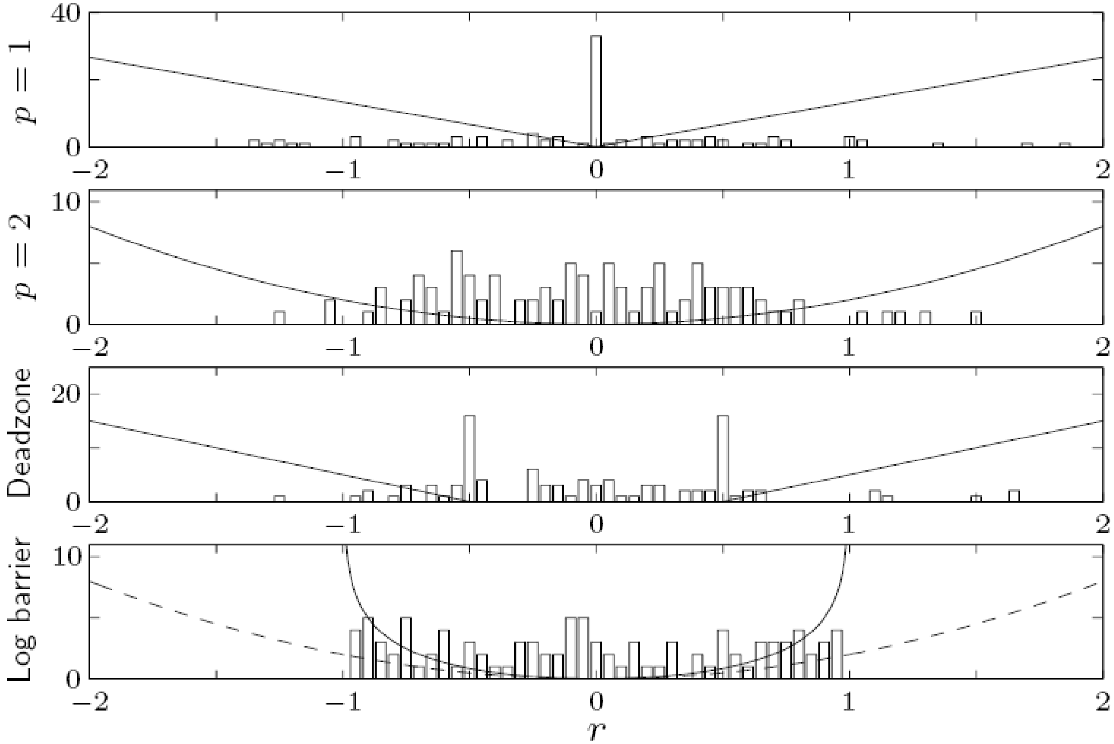


What you can recall?

8.1. Norm Approximation

8.1.2 Nonlinear Norm Approximation

For example, histogram of residuals for penalties $\phi(u) = |u|$, $\phi(u) = u^2$, $\phi(u) = \max\{0, |u| - a\}$, $\phi(u) = -\log(1 - u^2)$ are



8.1. Norm Approximation

8.1.2 Nonlinear Norm Approximation

The shape of penalty function has large effect on distribution of residuals.

Huber penalty function with parameter M

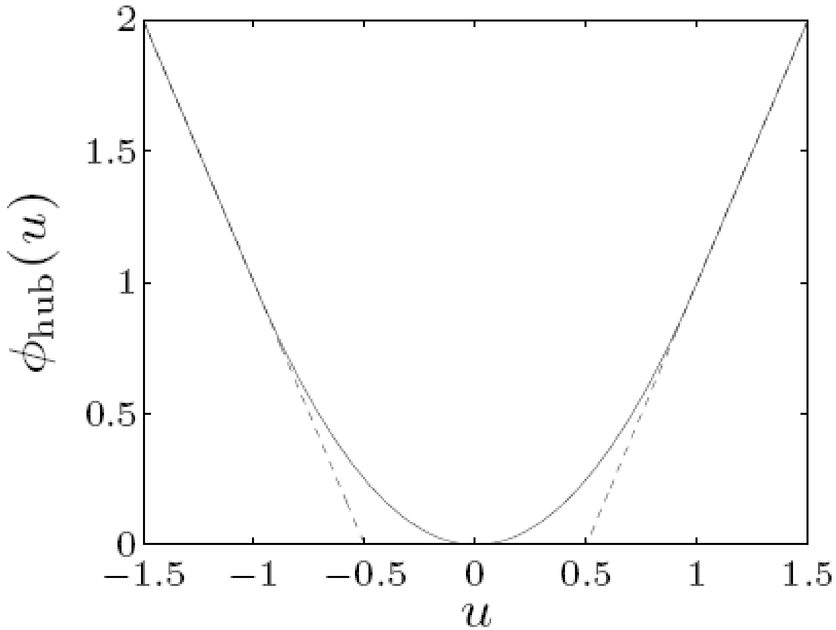
$$\phi_{hub}(u) = \begin{cases} u^2 & |u| < M \\ M(2|u| - M) & |u| \geq M \end{cases} \tag{8.7}$$

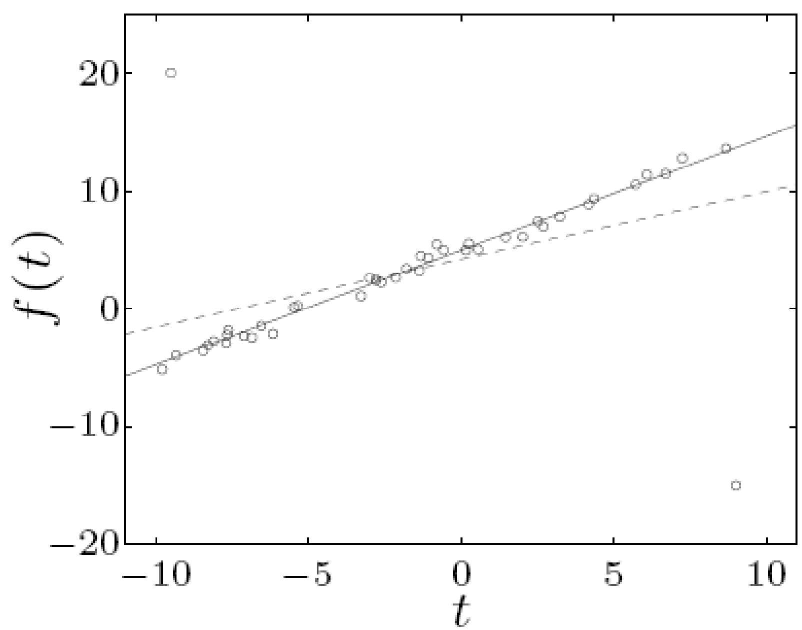
Linear growth for large u gives approximation less sensitive to outliners

8.1. Norm Approximation

8.1.2 Nonlinear Norm Approximation

left: Huber penalty for $M = 1$
right: affine function $f(t) = \alpha + \beta t$ fitted to 42 points t_i, y_j (circles), using quadratic (dashed) and Huber (solid) penalty





8.1. Norm Approximation

8.1.2 Nonlinear Norm Approximation

$$\text{minimize} \quad |Ax - b| \quad \text{with uncertain} \quad A \tag{8.8}$$

two approaches:

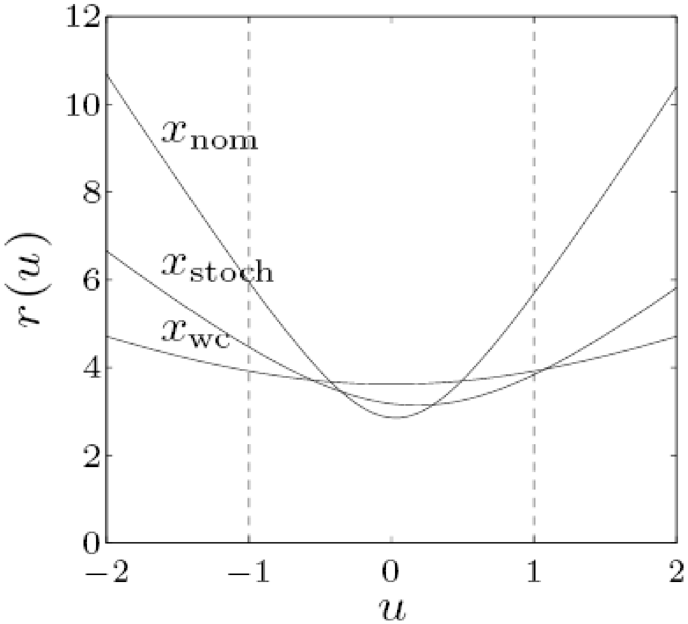
stochastic: assume A is random, minimize $E|Ax - b|$

worst-case: set A of possible values of A, minimize $\sup_{A \in Set(A)} |Ax - b|$, tractable only in special cases (certain norms $|\cdot|$, distributions, set of A)

8.1. Norm Approximation

8.1.2 Nonlinear Norm Approximation

example: $A(u) = A_0 + A_1$: x_{nom} minimizes $\|A_0 x - b\|_2^2$, x_{stoch} minimizes $E\|A(u)x - b\|_2^2$ with u uniform on $[-1,+1]$, x_{wc} minimizes $\sup_{-1 \le u \le 1} \|A(u)x - b\|_2^2$



8.1. Norm Approximation

8.1.2 Nonlinear Norm Approximation

stochastic robust LS with $A = \bar{A} + U$, U random $EU = 0$, $EU^T U = P$

$$\text{minimize} \qquad E\left\|(\bar{A} + U)x - b\right\|_2^2 \qquad (8.9)$$

Explicit expression for objective

$$E\left\|Ax - b\right\|_2^2 = E\left\|\bar{A}x - b + Ux\right\|_2^2 = \left\|\bar{A}x - b\right\|_2^2 + Ex^T U^T Ux \qquad (8.10)$$

where we can let $Ex^T U^T Ux = x^T Px$.

8.1. Norm Approximation

8.1.2 Nonlinear Norm Approximation

Thus, robust LS problem is equivalent to LS problem

$$\text{minimize} \quad \left|\bar{A}x - b\right|_2^2 + \left|P^{1/2}x\right|_2^2 \qquad (8.11)$$

It is indeed a Tikhonov regularization

worst-case robust LS with $A = \{\bar{A} + u_1 A_1 + \cdots + u_p A_p \mid |u|_2 \leq 1\}$

$$\text{minimize} \quad \sup_{A\in A} \left|Ax - b\right|_2^2 = \sup_{|u|_2 \leq 1} \left|P(x)u + q(x)\right|_2^2 \qquad (8.12)$$

where $P(x) = \begin{bmatrix} A_1 x & A_2 x & \cdots & A_p x \end{bmatrix}$, $q(x) = \bar{A}x - b$

8.1. Norm Approximation

8.1.2 Nonlinear Norm Approximation

strong duality holds between the following problems

$$
\begin{array}{llll}
\text{maximize} & \|Pu+q\|_2^2 & \text{minimize} & t+\lambda \\
\text{subject to} & \|u\|_2^2 \le 1 & \text{subject to} & \begin{bmatrix} I & P & q \\ P^T & \lambda I & 0 \\ q^T & 0 & t \end{bmatrix} \ge 0
\end{array}
\qquad (8.13)
$$

Indeed, it is somewhat like to find an ellipsoid with maximum volume.

8.1. Norm Approximation

8.1.2 Nonlinear Norm Approximation

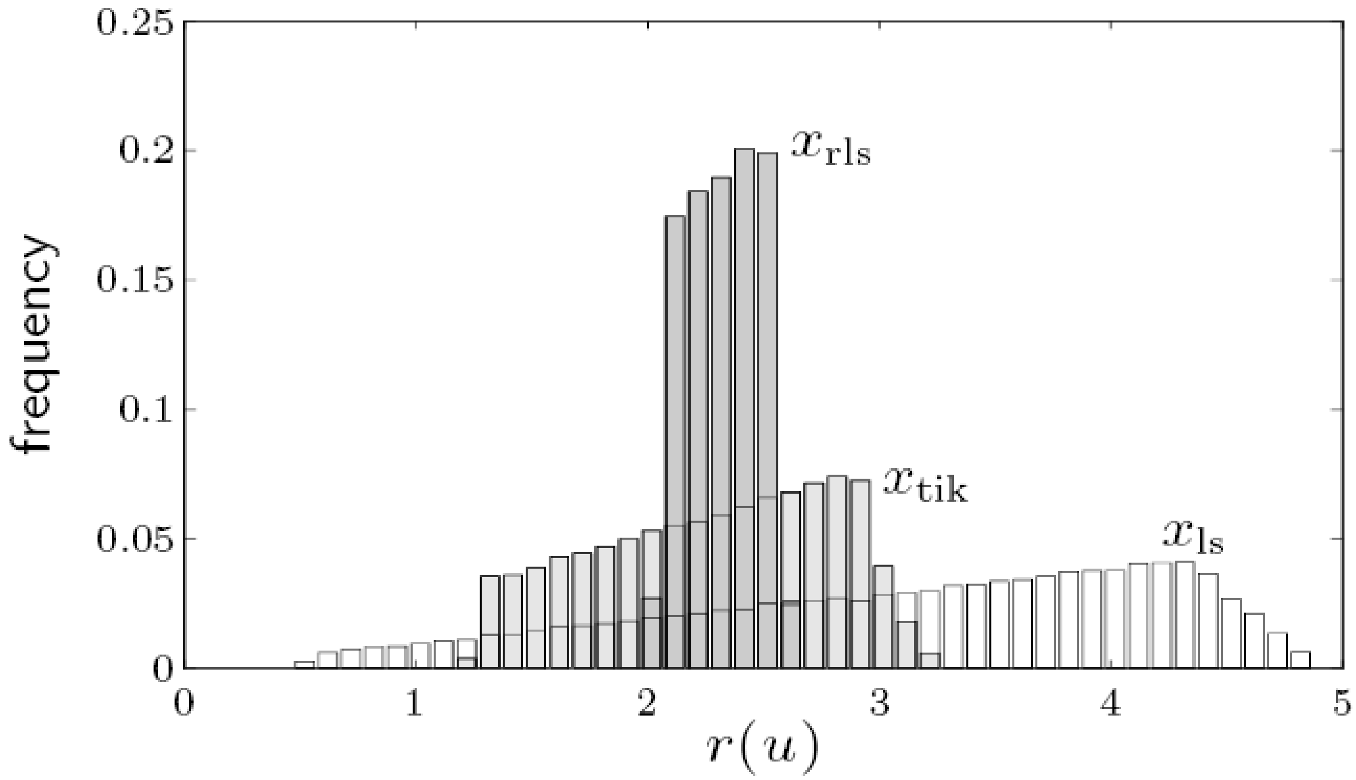
Hence, robust LS problem is equivalent to SDP

$$\begin{array}{ll} \text{minimize} & t+\lambda \\ \text{subject to} & \begin{bmatrix} I & P(x) & q(x) \\ P(x)^T & \lambda I & 0 \\ q(x)^T & 0 & t \end{bmatrix} \geq 0 \end{array} \qquad (8.14)$$

8.1. Norm Approximation

8.1.2 Nonlinear Norm Approximation

histogram of residuals $r(u) = \left|(A_0 + u_1 A_1 + u_2 A_2)x - b\right|_2$ with u uniformly distributed on unit disk, for three values of x



8.2. l_1 Magic

8.2.1 l_0 Problems and Sparsity

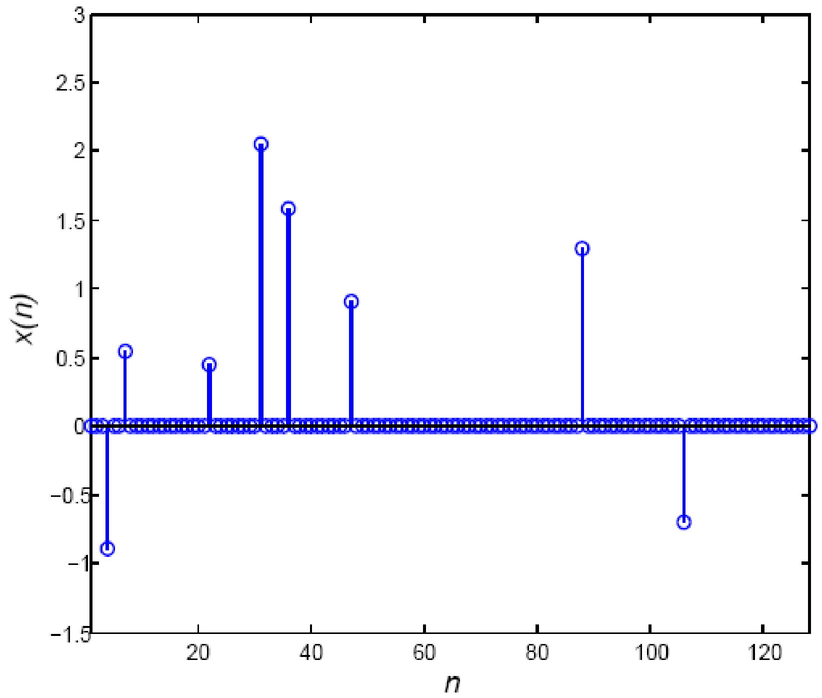
Suppose we need to reconstruct signal x from measurement $y = Ax$, where $A \in R^{m \times n}$, $x \in R^n$, $y \in R^m$, $m << n$.

In general this is not possible, but if x is K-sparse, meaning that it has only K nonzero entries then it is possible to design that preserve the information of x using only $m = O\left(K \log \frac{n}{K}\right)$ measurements. The most commonly studied A that satisfy this bound on M are random matrice, i.e., each entry of A is drawn independently from some suitable distribution.

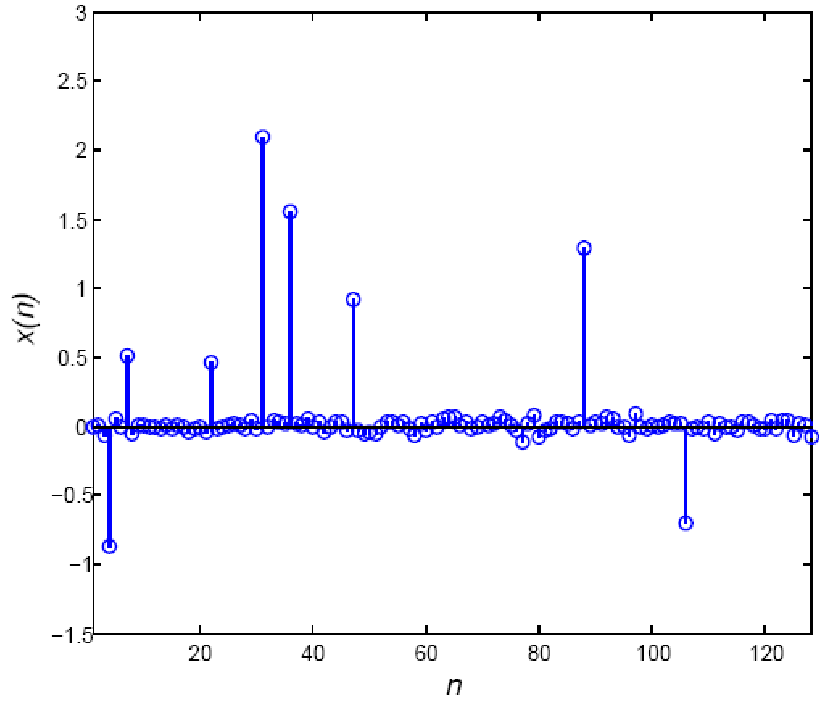
We really have no time to discuss Random Matrix Theory.

8.2. l_1 Magic

8.2.1 l_0 Problems and Sparsity



A sparse signal



A near sparse signal

A signal is near K -sparse if contains K significant components.

8.2. l_1 Magic

8.2.1 l_0 Problems and Sparsity

In general, recovering x from measurement vector $y = Ax$ with $m << n$ is an ill-posed problem. But if $Spark(A)$, the size of the smallest linearly dependent subset of A, satisfies that $Spark(A) \geq |x|_0$, we can accurately recover x by solving

$$\min |x|_0 \text{ subjecto to } y = Ax \tag{8.15}$$

where $|x|_0$ is simply the number of nonzero components in x, which is also known as l_0 norm of x.

But directly attack is NP hard!

8.2. l_1 Magic

8.2.2 l_1 Magic

If x is sparse enough, we can solve the following instead (with very high probability to hit the solution of (8.15))

$$\min |x|_1 \text{ subjecto to } y = Ax \tag{8.16}$$

Yes, we can transform it to an epigraph problem

$$\min 1^T t \text{ subjecto to } Ax = y,\ -t \le y \le t \tag{8.17}$$

This is called "*Decoding by linear programming*".

8.2. l_1 Magic

8.2.2 l_1 Magic

After getting $\hat{x}$ from (8.17), we can use thresholding method to choose the selected entries of x, and find x by resolving a well-defined linear equation set with reduced dimensions

$$\widetilde{A}\widetilde{x} = \widetilde{y} \qquad (8.18)$$

where $\widetilde{A} \in R^{m\times m}$, $\widetilde{x} \in R^{m}$, $\widetilde{y} \in R^{m}$.

There are some other approaches to get the real x through $\hat{x}$ obtained from (8.17). But we will not discuss them here.

8.2. l_1 Magic

8.2.2 l_1 Magic

Someone may ask why we do not solve

$$\min|x|_2 \text{ subjecto to } y = Ax \tag{8.19}$$

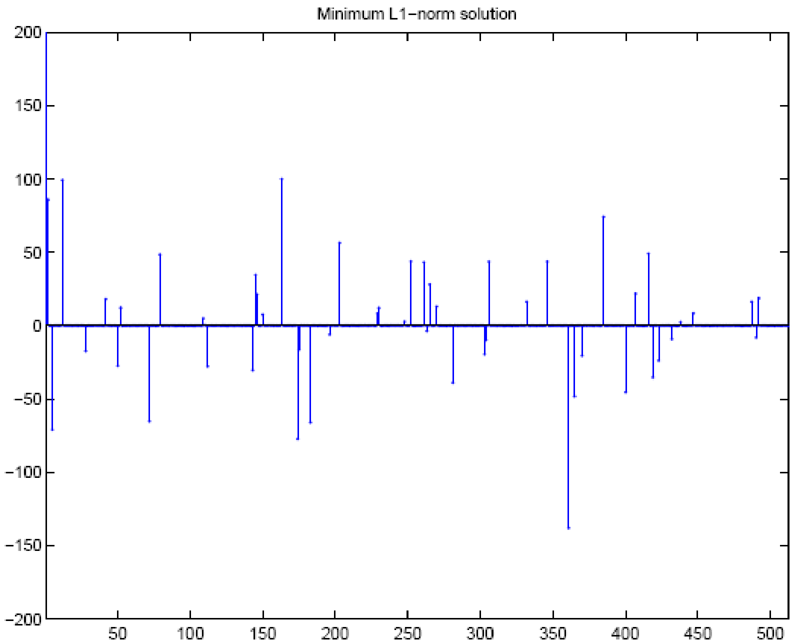
though we can solve it via LMS (Do you remember (8.3)?).

Actually, the idea of using the 1-norm instead of the 2-norm for better data recovery has been explored since mid-seventies in various applied areas, in particular geophysics and statistics. People find that the solution of (8.19) has low probability to hit the solution of (8.15).

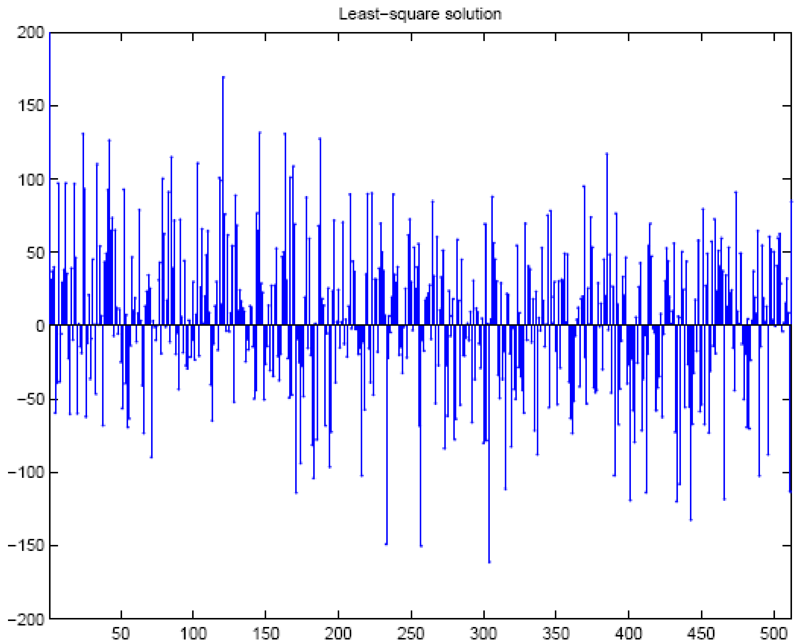
8.2. l_1 Magic

8.2.2 l_1 Magic

Minimum L_1-norm solution versus minimum L_2-norm (least-squares) solution:



L_1 solution



L_2 solution

8.2. l_1 Magic

8.2.2 l_1 Magic

```
m=256; n=512;
A = rand(m,n); x = zeros(n,1);
x(n/8,1) = 1;x(n/4,1) = -1;x(n/2, 1) = -1.5;b = A*x;

cvx_begin
   variable x_l1(n);
   minimize (norm( x_l1, 1)) % minimize (norm( x_l1, 2))
   subject to
      A*x_l1 == b;
cvx_end

plot(x_l1)
```

8.2. l_1 Magic

8.2.2 l_1 Magic

The magic of l_1 is that it combines the parsimony of l_0 and the computational efficiency of l_2.

The potential of using the 1-norm (which is also called Basis Pursuit) for exact reconstruction is illustrated by the following heuristics.

The minimizer u to (P02) is the contact point where the smallest Euclidean ball centered at y' meets the subspace Y. That contact point is in general different from y. The situation is much better in (P01): typically the solution coincides with y.

8.2. l_1 Magic

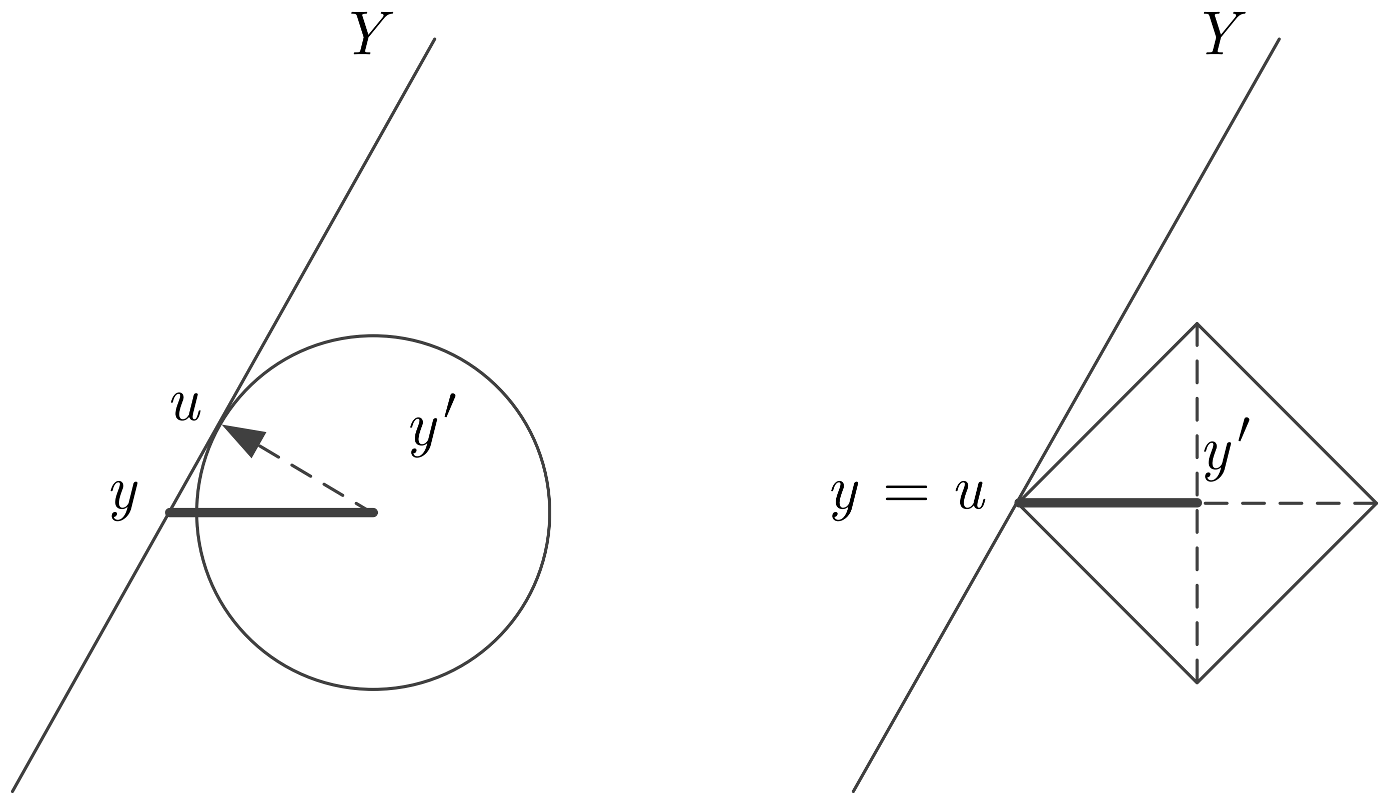
8.2.2 l_1 Magic

The minimizer u to (P01) is the contact point where the smallest octahedron centered at y' meets Y. Because the vector $y-y'$ lies in a low-dimensional coordinate subspace, the octahedron has a wedge at y.

Therefore, many subspaces Y through y will miss the octahedron of radius $y-y'$ (as opposed to the Euclidean ball). This forces the solution u to (P01), which is the contact point of the octahedron, to coincide with y.

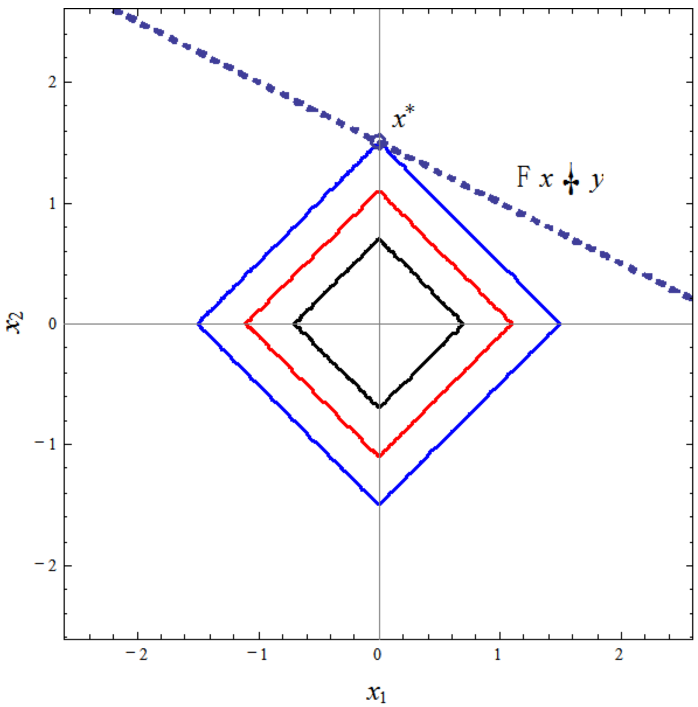
8.2. l_1 Magic

8.2.2 l_1 Magic



8.2. l_1 Magic

8.2.2 l_1 Magic

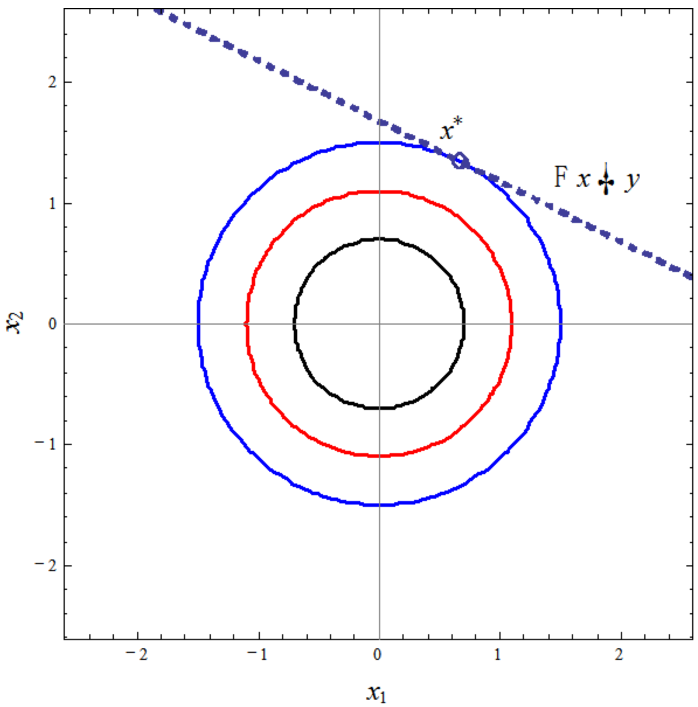


The contours for $|x|_1 = c$

As c increases, the contour grows and finally touches the hyperplane $Fx = y$, yielding a sparse solution $x = [0 \quad c]^T$.

8.2. l_1 Magic

8.2.2 l_1 Magic



The contours for $|x|_2 = c$

As c increases, the contour grows and finally touches the hyperplane $Fx = y$, yielding a non-sparse solution.

8.2. l_1 Magic

8.2.2 l_1 Magic

There are many variations

1)

$$\min |x|_1 \text{ subject to } |Ax - y|_p \le \varepsilon \tag{8.20}$$

which requires subgradient-based nonsmooth optimization

2)

$$\min |x|_1 + \lambda |Ax - y|_p \tag{8.21}$$

which is quasi LASSO (when $p = 2$, it is LASSO)

8.2. l_1 Magic

8.2.2 l_1 Magic

3)

$$\min \left| Ax - y \right|_p \text{ subjecto to } \left| x \right|_1 \le z \tag{8.22}$$

which is a kind of sparse learning.

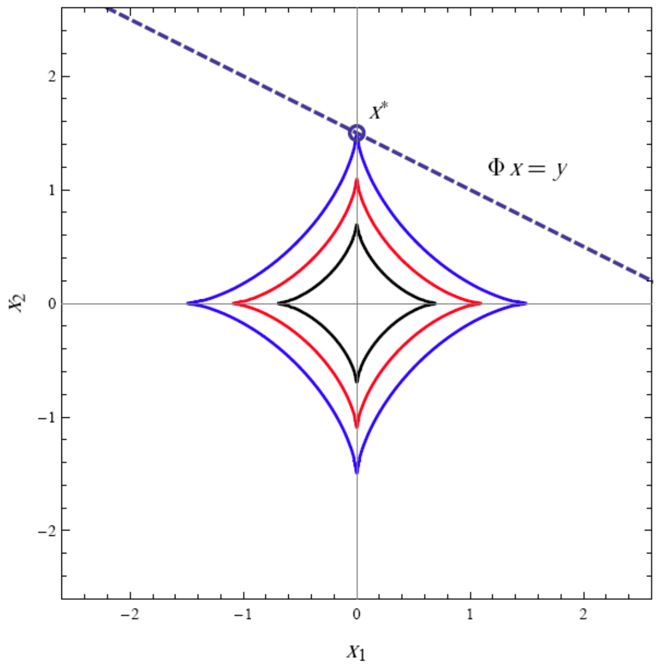
4)

$$\min \mu \left| x \right|_p + \lambda \left| Ax - y \right|_q \tag{8.23}$$

which can be more general cases, especially p and q are not integer! Usually, we set p, $q < 1$.

8.2. l_1 Magic

8.2.2 l_1 Magic



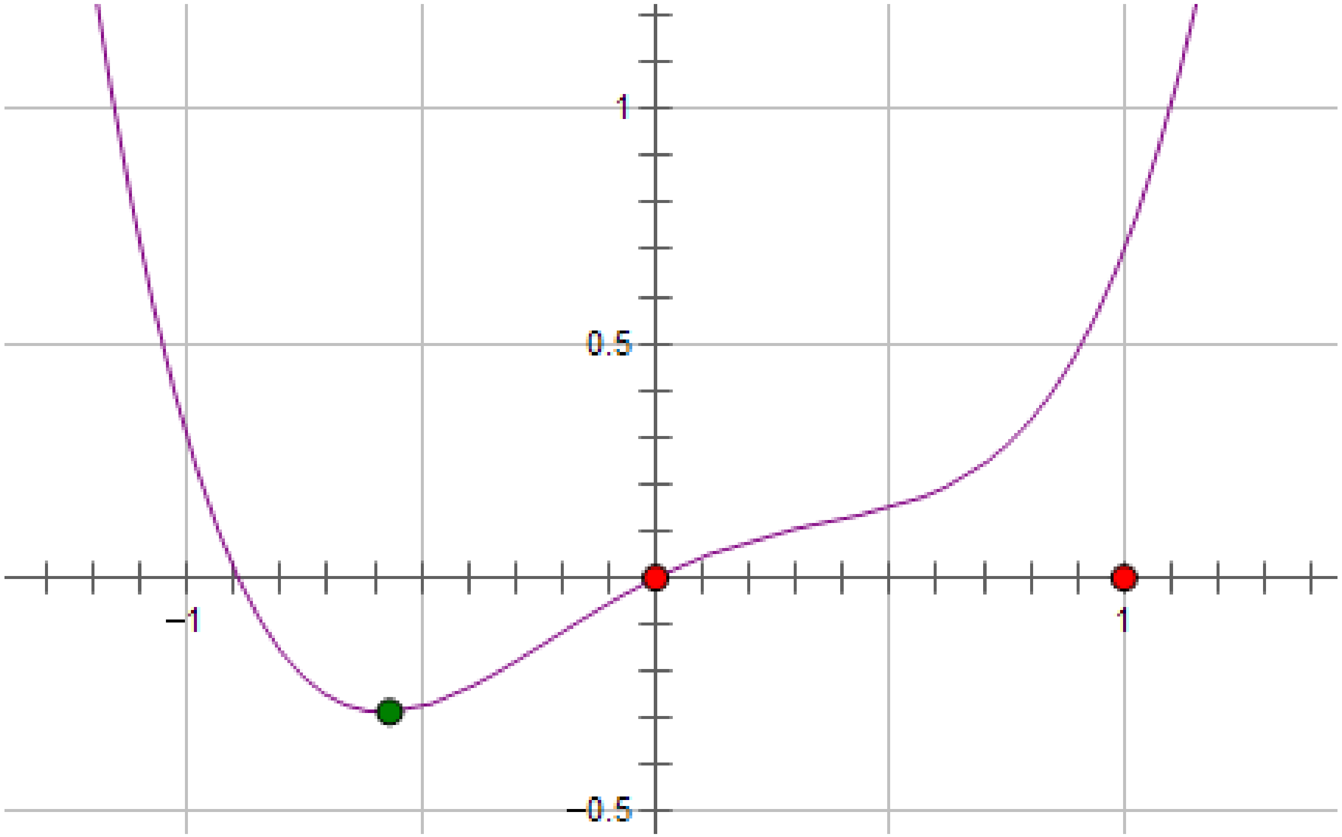
The contours for $|x|_p = c$, $p < 1$

As c increases, the contour grows and finally touches the hyperplane $Fx = y$, yielding a sparse solution $x = [0 \quad c]^T$.

8.2. l_1 Magic

8.2.2 l_1 Magic

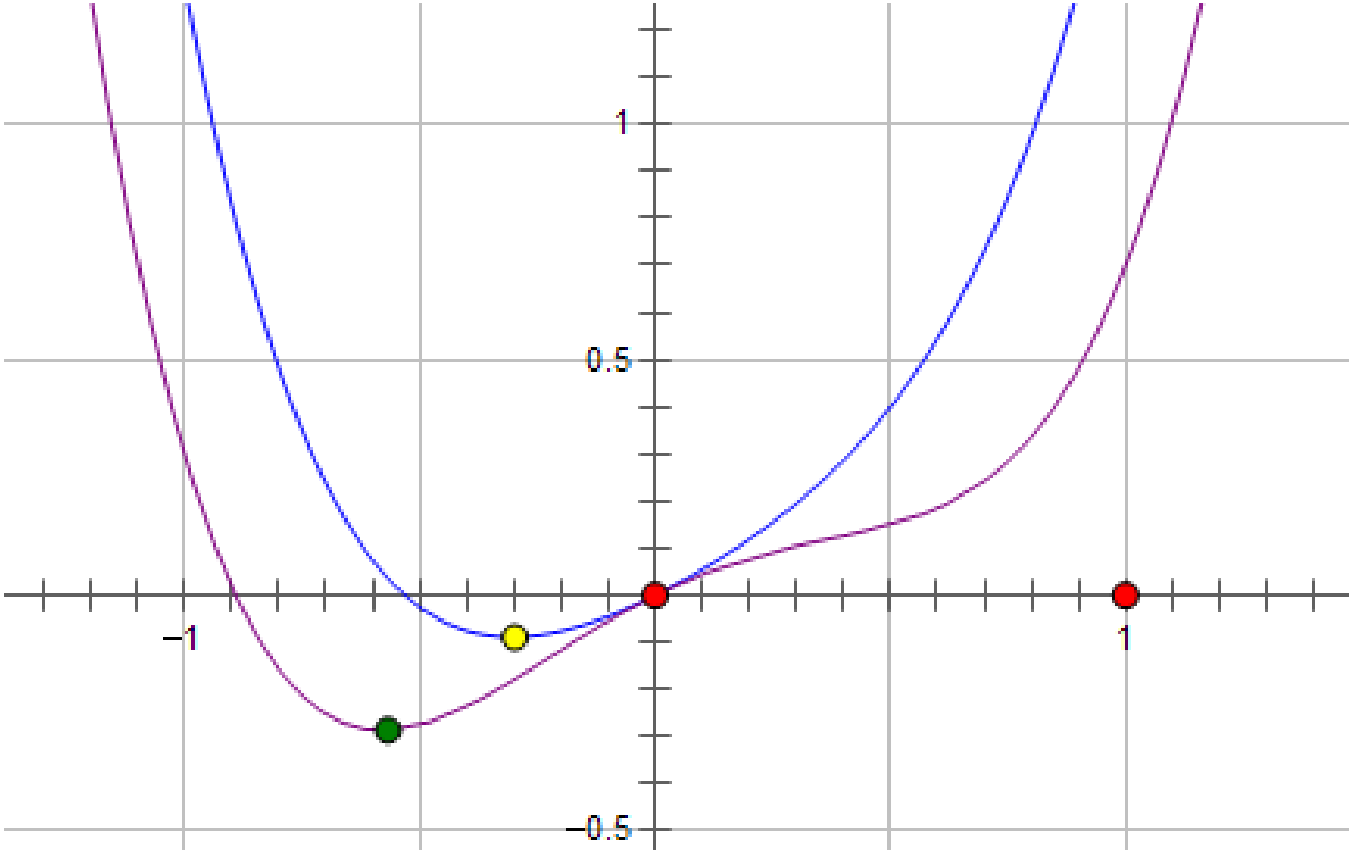
原函数 L 最优的 x 在绿点处，x 非零



8.2. l_1 Magic

8.2.2 l_1 Magic

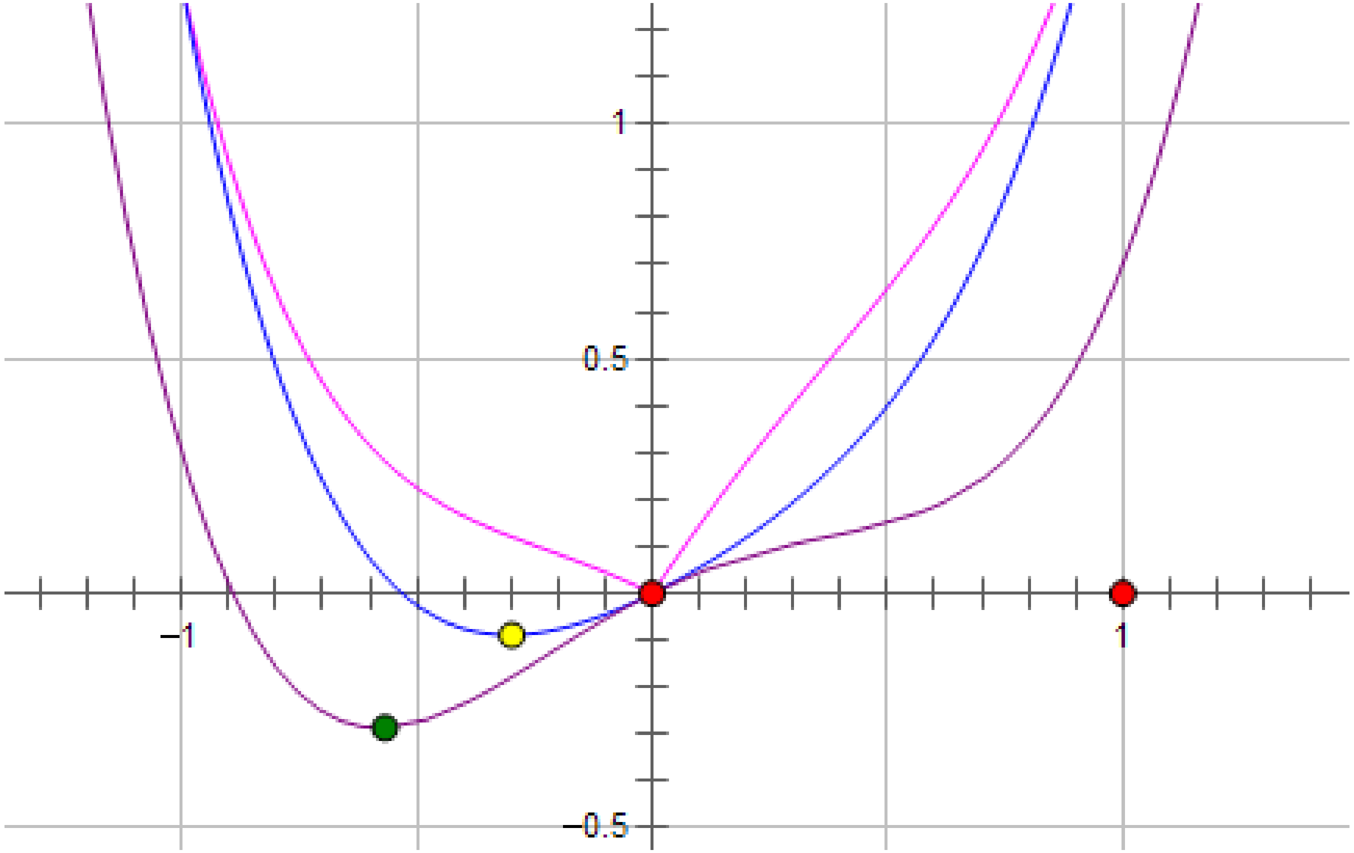
施加 L2 regularization 得到新的函数，最优的 x 在黄点处，x 的绝对值减小了，但依然非零



8.2. l_1 Magic

8.2.2 l_1 Magic

施加 L1 regularization 得到新的函数，最优的 x 就变成了 0。这里利用的就是绝对值函数的尖峰



8.2. l_1 Magic

8.2.2 l_1 Magic

一个不那么普适的观察结果

两种 regularization 能不能把最优的 x 变成 0，取决于原先的费用函数在 0 点处的导数。如果本来导数不为 0，那么施加 L2 regularization 后导数依然不为 0，最优的 x 也不会变成 0。而施加 L1 regularization 时，只要 regularization 项的系数 C 大于原先费用函数在 0 点处的导数的绝对值，$x = 0$ 就会变成一个极小值点。

作者：王赟 Maigo

https://www.zhihu.com/question/37096933/answer/70426653

8.2. l_1 Magic

8.2.2 l_1 Magic

5) If we need to reconstruct signal x from measurement $y = Ax + z$, where $A \in R^{m\times n}$, $m << n$, z are iid $N(0,\sigma^2)$. We can solve the follows with very high probability to recover x

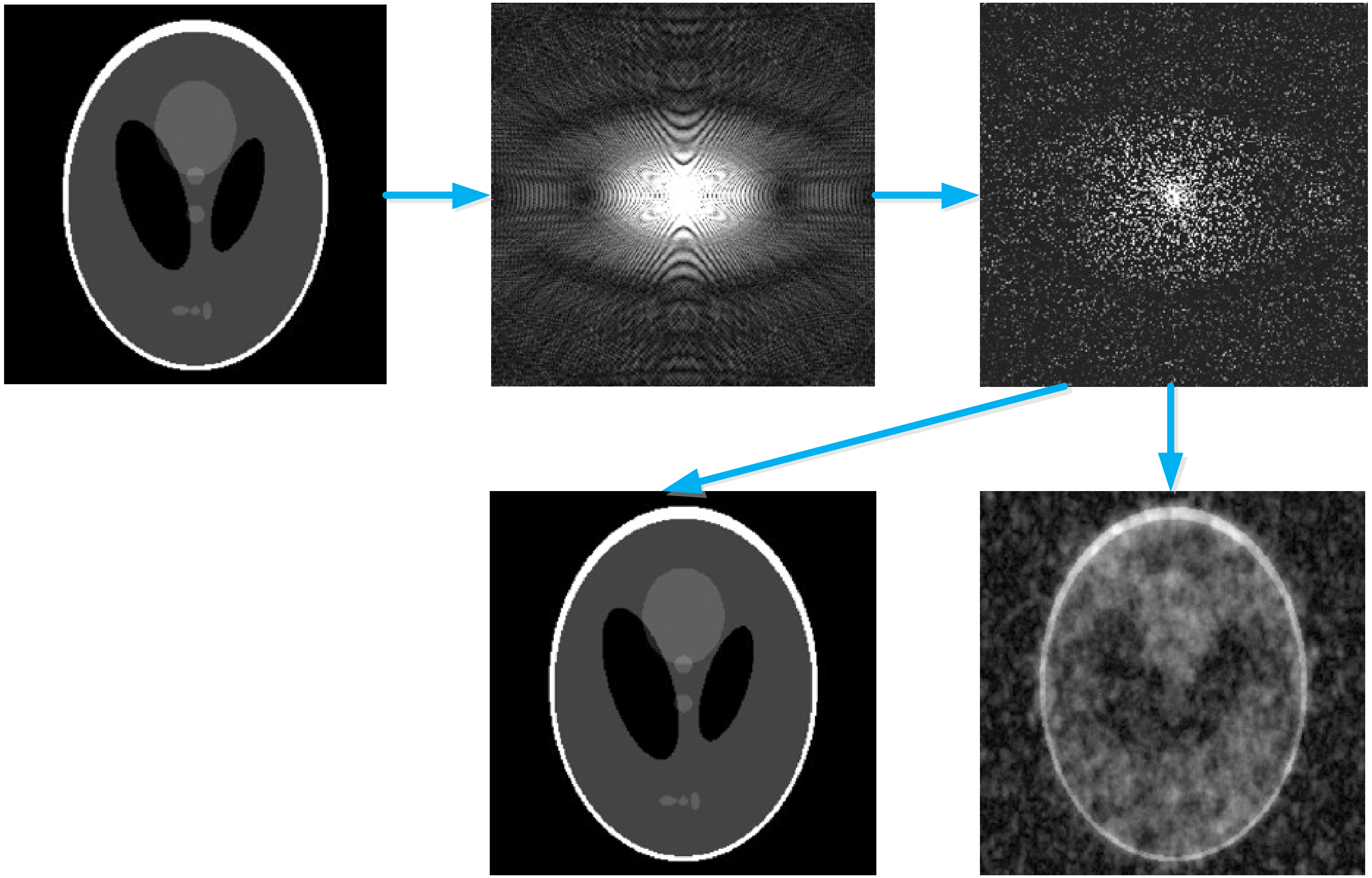
$$\min |x|_1 \tag{8.24}$$

subject to $\left|A^T(y - Ax)\right|_\infty \le (1 + t^{-1})\sigma\sqrt{2\log n}$, and t is a positive scalar. This is called "*Dantzig Selector*".

Currently, there is not common understanding on which one is better yet. So, this is a promising research direction.

8.2. l_1 Magic

8.2.3 Compressive Sensing



Recover after randomly throw 84% of samples for FT (from Prof. Lustig)

8.3. Regularization

8.3.1 Tikhonov Regularization

$$\text{minimize (w.r.t. } R_+^2\text{)} \ (|Ax - b|, |x|) \qquad (8.25)$$

$A \in R^{m \times n}$, norms on R^m and R^n can be different

Interpretation: find good approximation $Ax \approx b$ with small x

Estimation: linear measurement model $y = Ax + v$, with prior knowledge that $|x|$ is small

Optimal design: small x is cheaper or more efficient, or the linear model $y = Ax$ is only valid for small x

Robust approximation: good approximation $Ax \approx b$ with small x is less sensitive to errors in A than good approximation with large x

8.3. Regularization

8.3.1 Tikhonov Regularization

$$\text{minimize} \quad |Ax - b| + \gamma |x| \tag{8.26}$$

Tikhonov regularization:

$$\text{minimize} \quad |Ax - b|_2^2 + \delta |x|_2^2 \tag{8.27}$$

$$\text{minimize} \quad \left\| \begin{bmatrix} A \\ \sqrt{\delta} I \end{bmatrix} x - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|_2^2 \tag{8.28}$$

The solution is $x^* = (A^T A + \delta I)^{-1} A^T b$

8.3. Regularization

8.3.1 Tikhonov Regularization

We can immediate to establish a link between the form of the loss and the solution properties, via the Tikhonov regularization framework.

Tikhonov regularization is a commonly used regularization method of ill-posed problems named for Andrey Tychonoff. In statistics, the method is also called ridge regression.

The motivation is: to stabilize the solution by adding some auxiliary nonnegativefunctional that embeds prior information about the solution.

8.3. Regularization

8.3.1 Tikhonov Regularization

The standard approach to solve an overdetemined system of linear equations give as

$$Ax = b \tag{8.29}$$

is known as linear least squares and seeks to minimize the residual

$$\left| Ax - b \right|^2 \tag{8.30}$$

where $\left| \cdot \right|$ is the Euclidean norm. However, the matrix A may be ill-conditional or singular yielding a large number of solutions.

8.3. Regularization

8.3.1 Tikhonov Regularization

In order to give preference to a particular solution with desirable properties, the regularization term is included in this minimization

$$\left|Ax - b\right|^2 = \left|\Gamma x\right|^2 \tag{8.31}$$

for some suitably chosen **Tikhonov matrix**, Γ. In many cases, this matrix is chosen as the identity matrix $\Gamma = I$, which aims to find solutions with smaller norms. In other cases, highpass operators (e.g., a difference operator or a weighted Fourier operator) may be used to enforce smoothness if the underlying vector is believed to be mostly continuous.

8.3. Regularization

8.3.1 Tikhonov Regularization

This regularization improves the conditioning of the problem, thus enabling a numerical solution. An explicit solution, denoted by $\hat{x}$ is given by

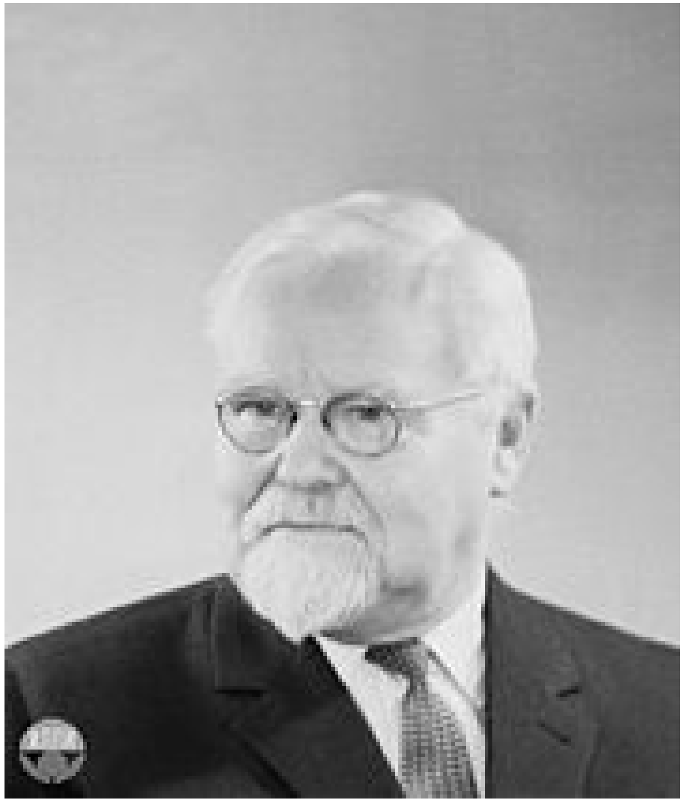
$$\hat{x} = (A^T A + \Gamma^T \Gamma)^{-1} A^T b \tag{8.32}$$

The effect of regularization may be varied via the scale of matrix Γ. For $\Gamma = \alpha I$, if $\alpha = 0$, it reduces to the unregularized least squares solution provided that $(A^T A)^{-1}$ exists.

Following Hoerl, Tikhonov Regularization is known as ridge regression in the statistical literature.

8.3. Regularization

8.3.1 Tikhonov Regularization



Andrey Nikolayevich Tychonoff
October 30, 1906 Gzhatsk - November 8, 1993 Moscow

8.3. Regularization

8.3.2 Regularization in Learning

Least-square measures

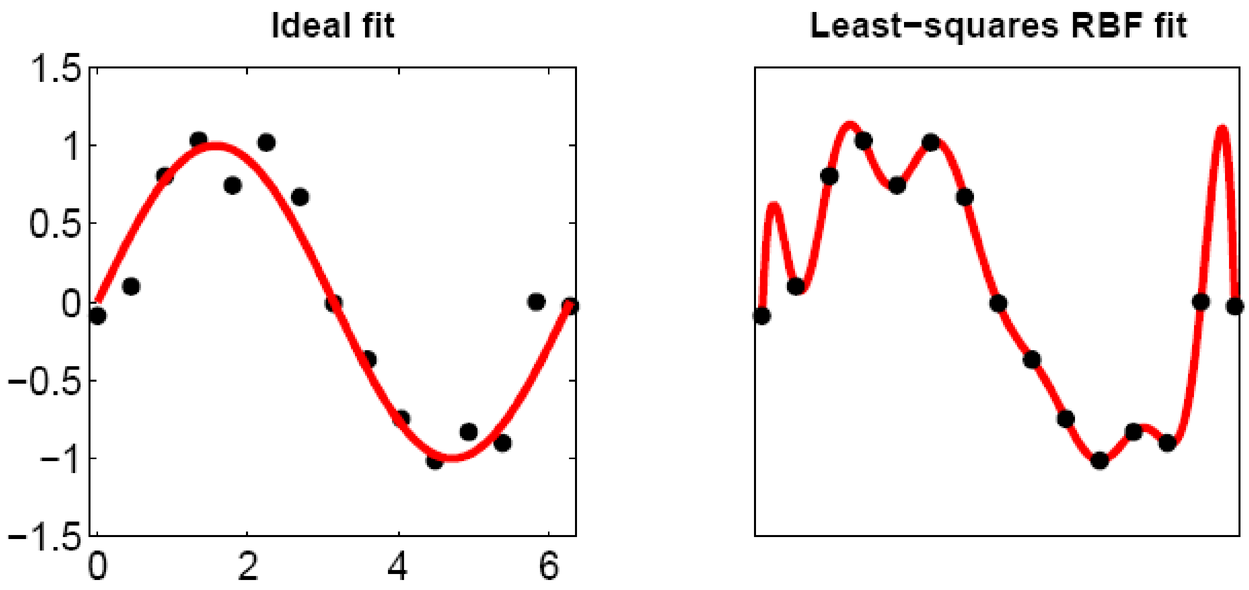
$$E_D(w) = \frac{1}{2}\sum_{n=1}^{N}\left[t_n - \sum_{m=1}^{M} w_m \phi_m(x_n)\right]^2 \qquad (8.33)$$

Let $t = [t_1, \ldots, t_N]^T$, $\Phi_{nm} = \phi_m(x_n)$ denote the design matrix, we have the estimator of w_m as

$$w_{LS} = \left[\Phi^T \Phi\right]^{-1} \Phi^T t \qquad (8.34)$$

8.3. Regularization

8.3.2 Regularization in Learning



The 'ideal' fit is shown on the left, while the least-squares fit using 15 basis functions is shown on the right and perfectly interpolates all the data points. But, …, Overfit?

8.3. Regularization

8.3.2 Regularization in Learning

Least-square+Regularization (Tikhonov Regularization)

$$E_D(w) + \lambda E_w(w) = \frac{1}{2}\sum_{n=1}^{N}\left[t_n - \sum_{m=1}^{M} w_m \phi_m(x_n)\right]^2 + \frac{\lambda}{2}\sum_{m=1}^{M} w_m^2 \qquad (8.35)$$

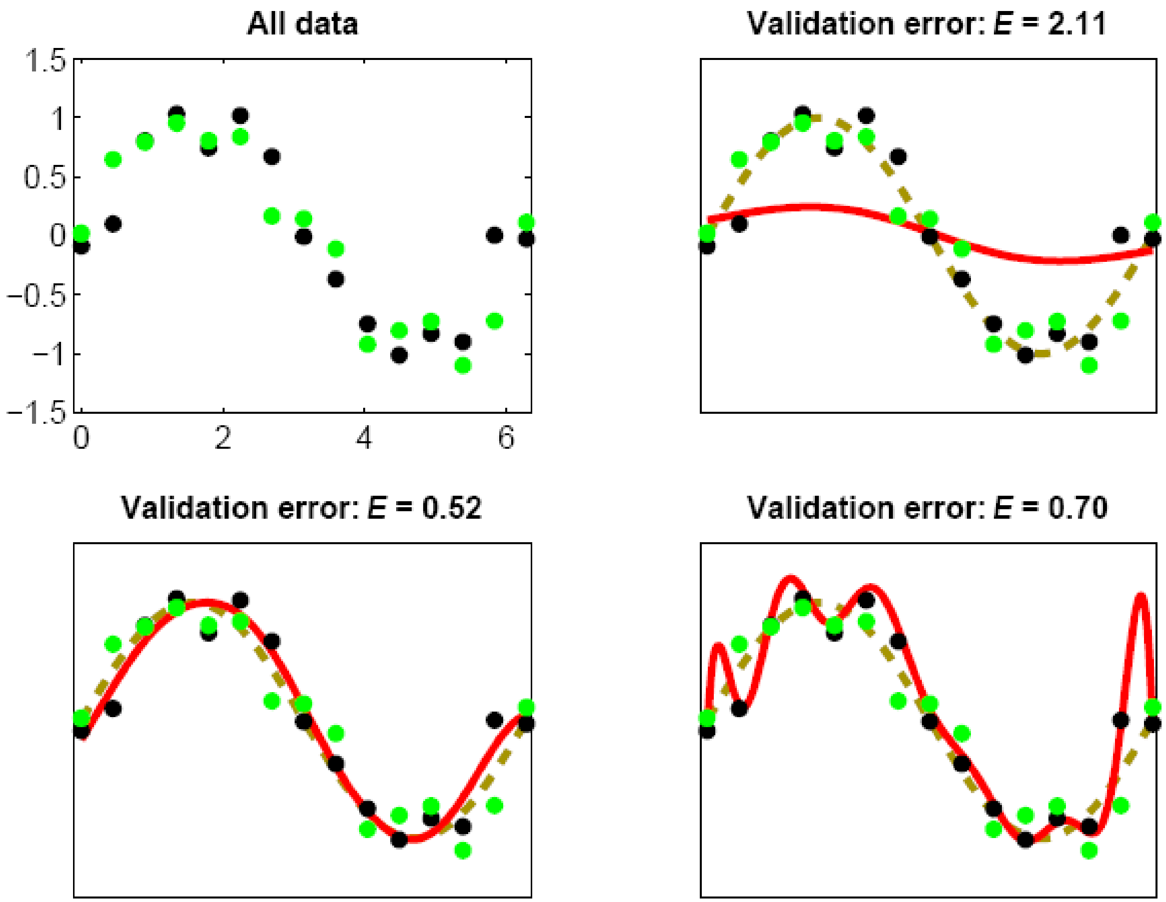
The penalized least-squares gives

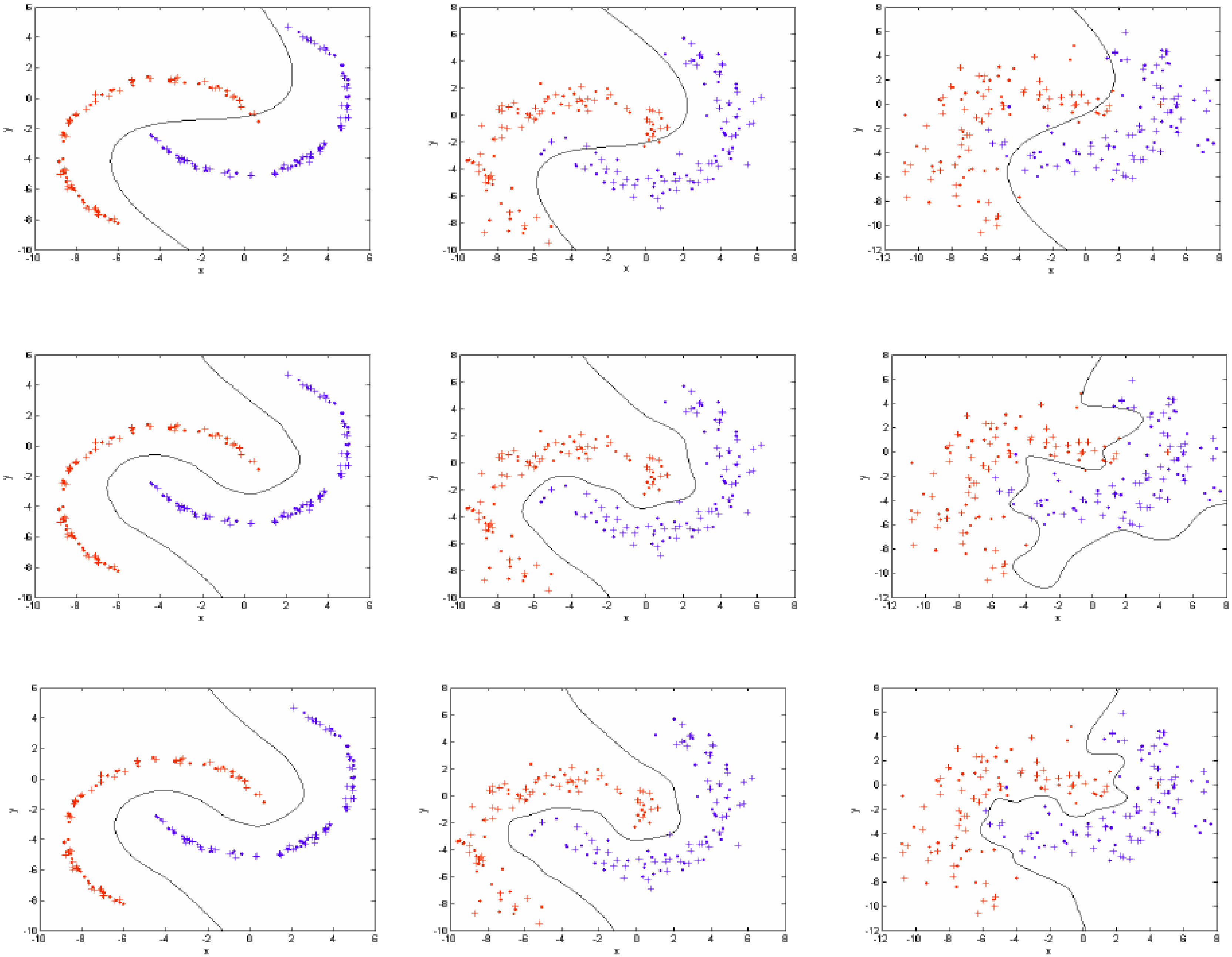
$$w_{PLS} = \left[\Phi^T \Phi + \lambda I\right]^{-1} \Phi^T t \qquad (8.36)$$

where the hyperparameter λ balances the trade-off between $E_D(w)$ and $E_w(w)$.

8.3. Regularization

8.3.2 Regularization in Learning





8.3. Regularization

8.3.2 Regularization in Learning

Let us consider a more general case

$$t_n = y(x_n, w) + \varepsilon_n \tag{8.37}$$

where ε_n follows normal distribution, the likelihood can then given as

$$p(t \mid x, w, \sigma^2) = \prod_{n=1}^{N} p(t_n \mid x_n, w, \sigma^2) = \prod_{n=1}^{N} \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left(-\frac{\left[t_n - y(x_n, w)\right]^2}{2\sigma^2}\right) \tag{8.38}$$

8.3. Regularization

8.3.2 Regularization in Learning

The likelihood is

$$-\log p(t \mid x, w, \sigma^2) = \frac{N}{2}\log\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left[t_n - y(x_n, w)\right]^2 \tag{8.39}$$

Since the first term on the right in (8.39) is independent of w, we only need to consider the second term which is proportional to the squared error.

It is conspicuously a LS problem.

8.3. Regularization

8.3.2 Regularization in Learning

Maximum likelihood estimation often suffers from overfitting, if we want to get a solution for w via least-squares. Thus, we usually minimize

$$\hat{E}(w) = -\log p(t \mid w, \sigma^2) - \log p(w \mid \alpha) \tag{8.40}$$

where $p(w \mid \alpha)$ is a *prior distribution* which expresses our "degree of belief" over values that w might take.

Here, α is some pre-selected hyperparameters. Now, the question is: how to choose a meaningful yet computation convenient $p(w \mid \alpha)$?

8.3. Regularization

8.3.2 Regularization in Learning

$$p(w \mid \alpha) = \prod_{m=1}^{M} \left(\frac{\alpha}{2\pi} \right)^{\frac{1}{2}} \exp\left(-\frac{\alpha}{2} w_m^2 \right) \tag{8.41}$$

Though the prior is independent for each weight, the shared inverse variance hyperparameter α (can be taken as a certain Lagrange penalty coefficient) moderates the strength of our "belief". Substitue (8.41) to (8.40), retaining only those terms dependent on w, we have

$$\tilde{E}(w) = \frac{1}{2\sigma^2} \sum_{n=1}^{N} \left[t_n - y(x_n, w) \right]^2 + \frac{\alpha}{2} \sum_{m=1}^{M} w_m^2 \tag{8.42}$$

8.3. Regularization

8.3.2 Regularization in Learning

So it is analogue to Tikhonov Regularization with $\lambda = \sigma^2 \alpha$. If we define

$$y(x_n, w) = \sum_{m=1}^{M} w_m \phi_m (x_n) \quad (8.43)$$

We have the Tikhonov Regularization type MLE solution as

$$\mu = \left[\Phi^T \Phi + \sigma^2 \alpha I\right]^{-1} \Phi^T t \quad (8.44)$$

$$\Sigma = \sigma^2 \left[\Phi^T \Phi + \sigma^2 \alpha I\right]^{-1} \quad (8.45)$$

8.3. Regularization

8.3.2 Regularization in Learning

Actually, we can explain (8.42) via the Bayes's rule, too

$$p(t \mid w, \sigma^2) = \frac{p(t \mid w, \sigma^2)\, p(w \mid \alpha)}{p(t \mid \alpha, \sigma^2)} \tag{8.46}$$

Since $p(t \mid \alpha, \sigma^2)$ is a constant, we can get (8.41) from (8.46) directly.

In other words, we had chosen the zero-mean Gaussian prior, which expresses a preference for smoother models by declaring smaller weights to be *a priori* more probable.

8.3. Regularization

8.3.2 Regularization in Learning

Similarly, if we assume

$$p(w \mid \alpha) = ? \tag{8.47}$$

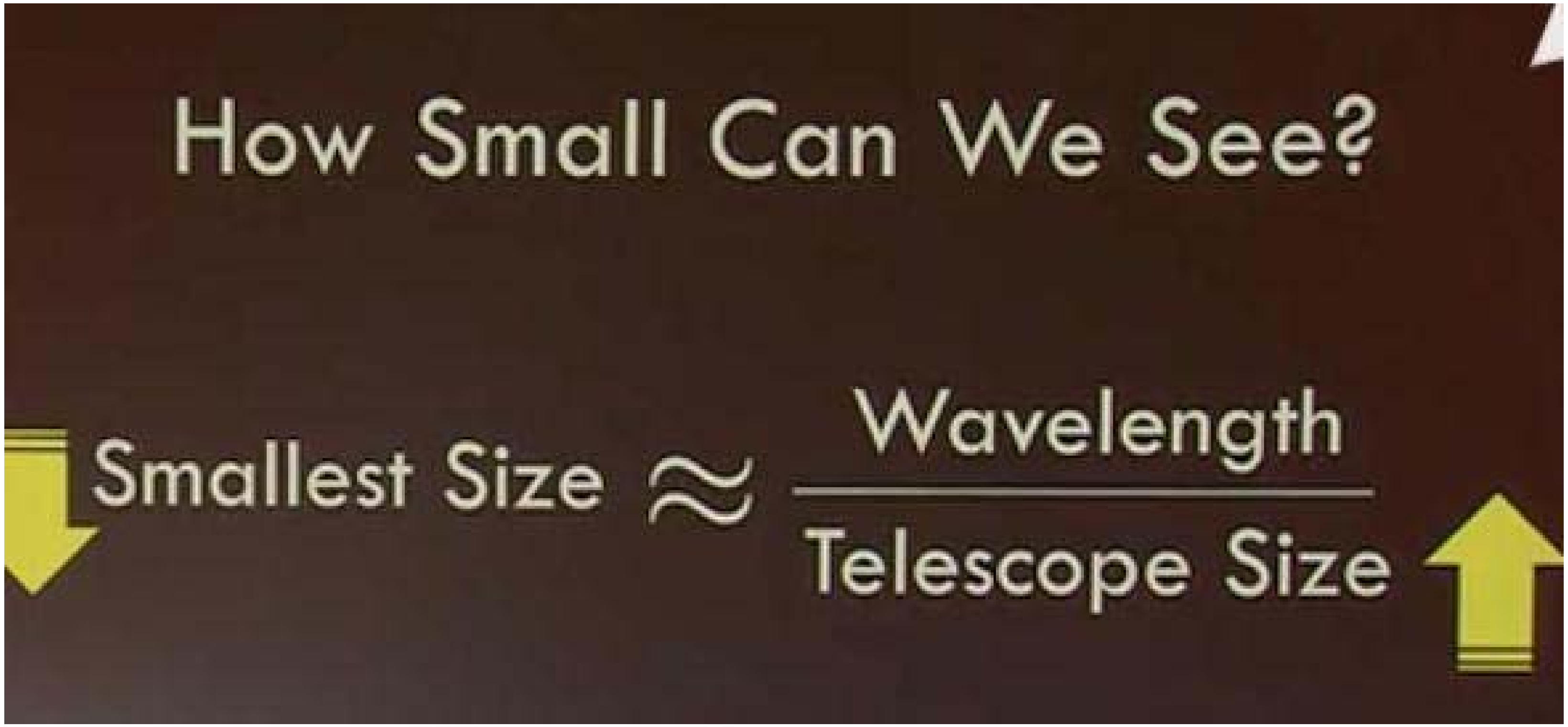
we can get the sparcity regularization

$$\widetilde{E}(w) = \frac{1}{2\sigma^2} \sum_{n=1}^{N} \left[t_n - y(x_n, w) \right]^2 + \frac{\alpha}{2} \sum_{m=1}^{M} \left| w_m \right| \tag{8.48}$$

8.3. Regularization

8.3.3 Imaging the Invisiable

衍射现象公式告诉我们：能看到的最小物体是有限的，你想看到的物体越小，所需的望远镜就需要越大。



8.3. Regularization

8.3.3 Imaging the Invisiable

这次拍摄的黑洞中第一个观测对象位于人马座方位，距离地球 2.6 万光年。第二个观测对象为室女座星系团中超大质量星系 Messier 87 中心的黑洞。该黑洞距离地球 5500 万光年，质量为太阳的 65 亿倍。要拍摄成功，需要地球大小的巨无霸光学望远镜才可以实现。而造这样一个望远镜，在人类目前的技术水平下，几乎是不可能的。

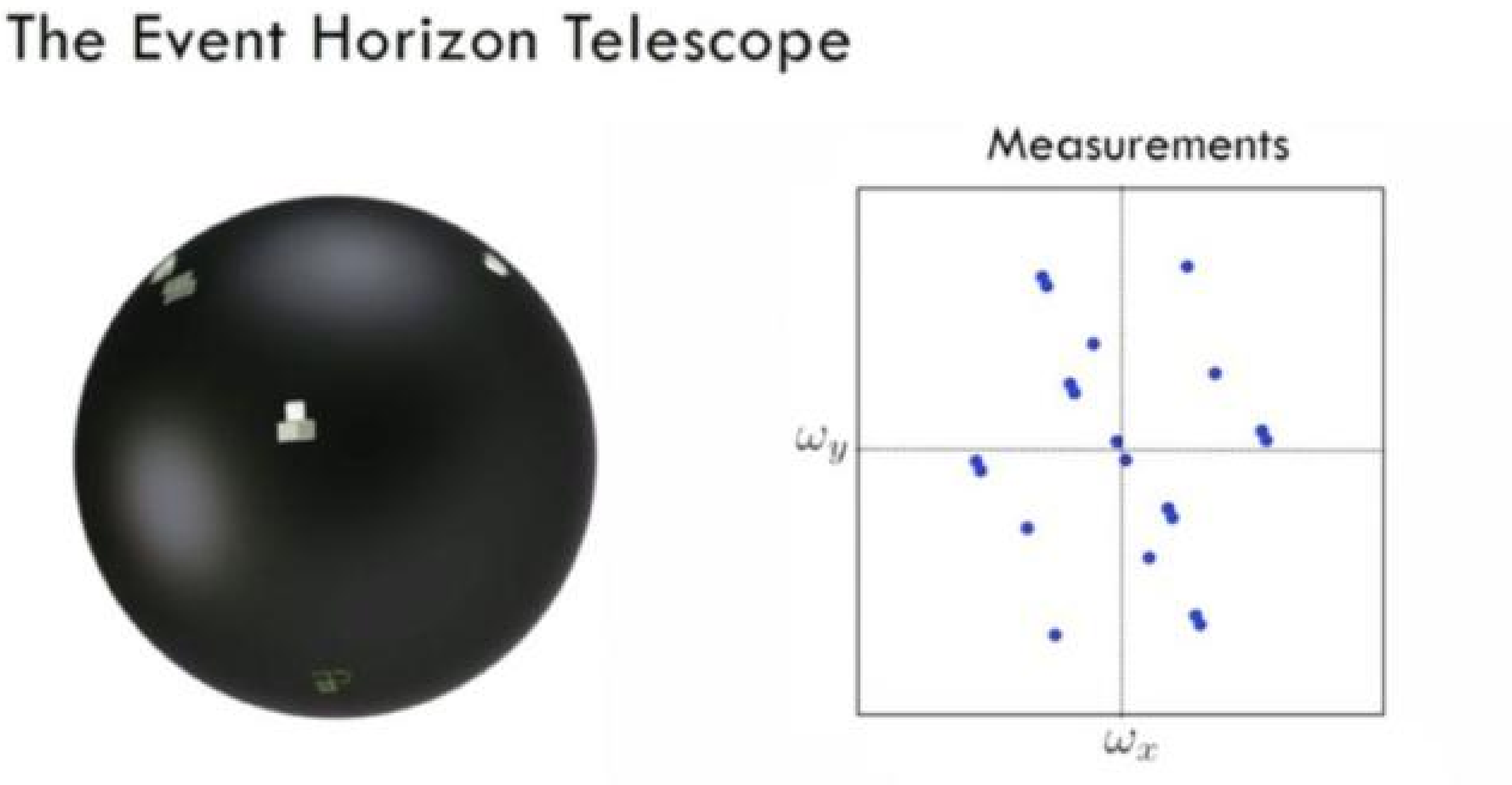
在探寻黑洞的路上，一个被称作“视界线望远镜”（Event Horizon Telescope）的团队提出了一个有点疯狂的想法：如果把世界上所有的望远镜连接起来，在电脑上模拟一个地球大小的望远镜，收集这些数据，再通过合理的算法分析，也许就能描绘黑洞的模样。

8.3. Regularization

8.3.3 Imaging the Invisiable

8.3. Regularization

8.3.3 Imaging the Invisiable

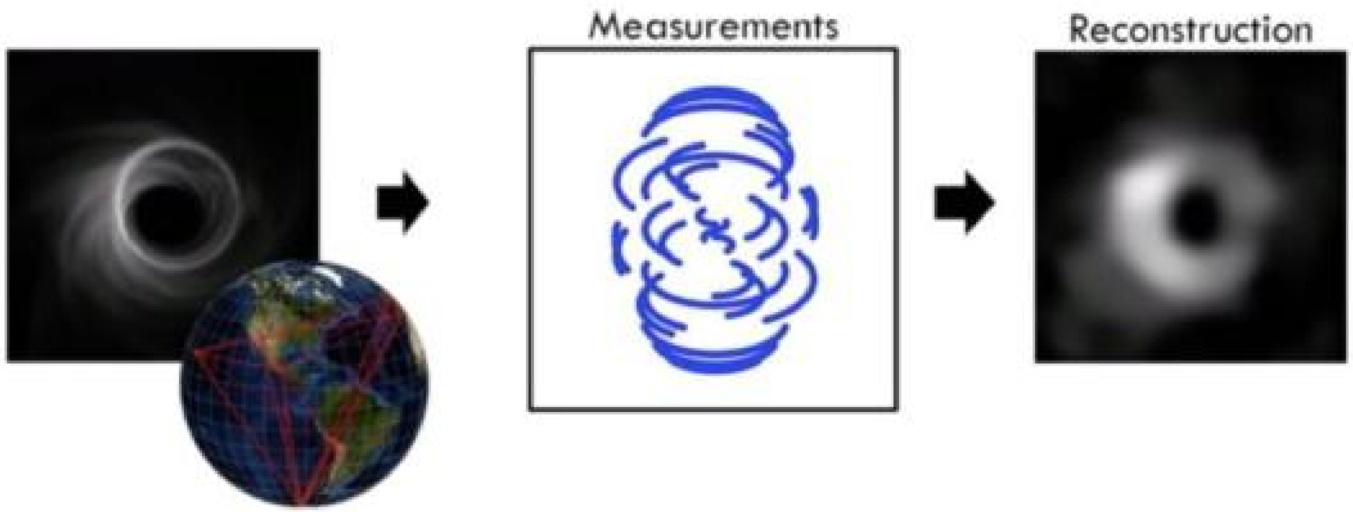


左边这黑球上的几个亮点就是我们现在布局在地球上的 8 个射电望远镜。通过他们收集的图像可平铺结合成右边的图。

8.3. Regularization

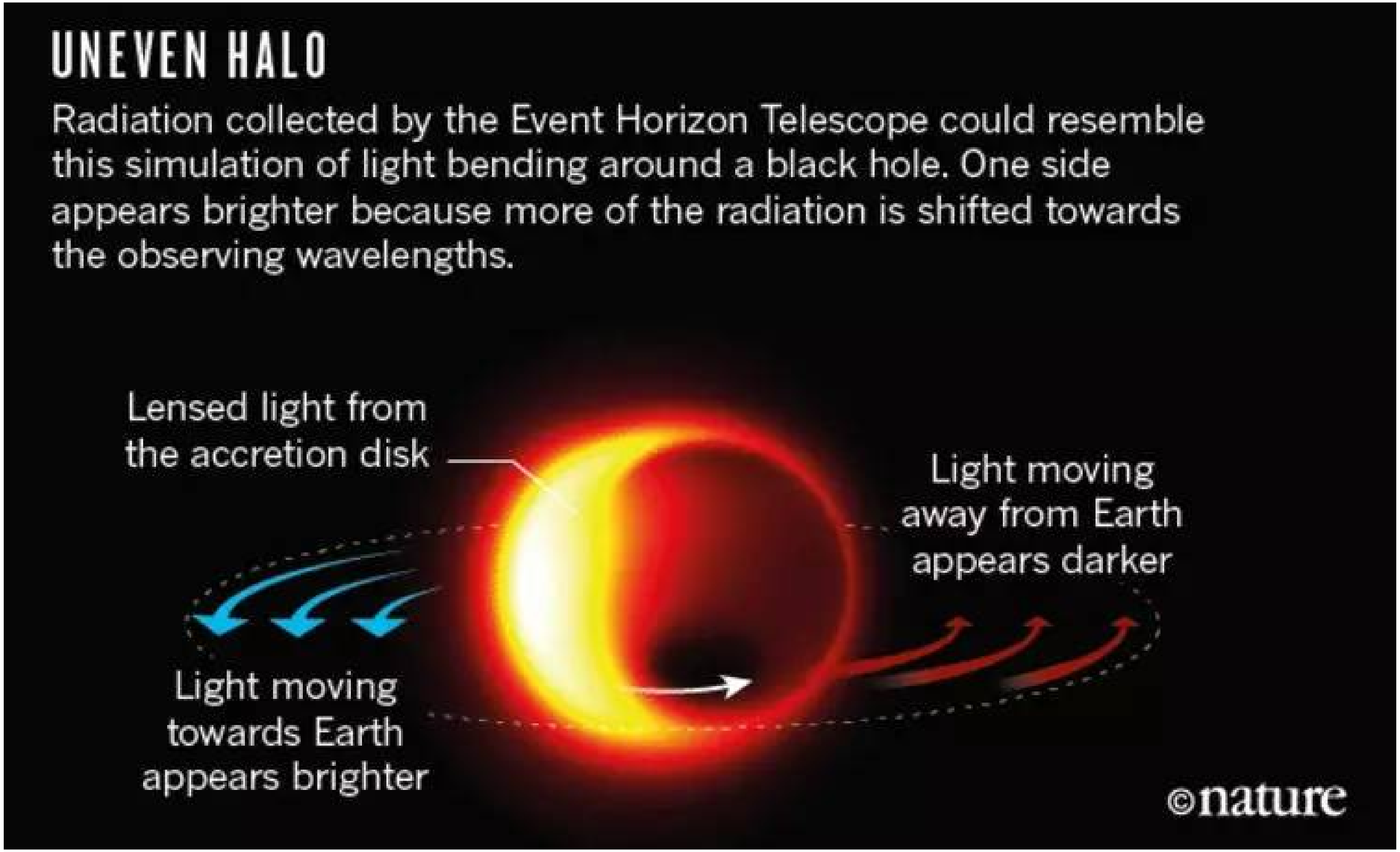
8.3.3 Imaging the Invisiable

虽然只有这可怜的几处，但随着地球的自转，望远镜的位置改变， 我们就可以得到图片的各个部分。但所得到的样本仍然不足，需要通过开发生成图片的算法将空白的地方填满， 从而模拟出隐藏的黑洞图片。



8.3. Regularization

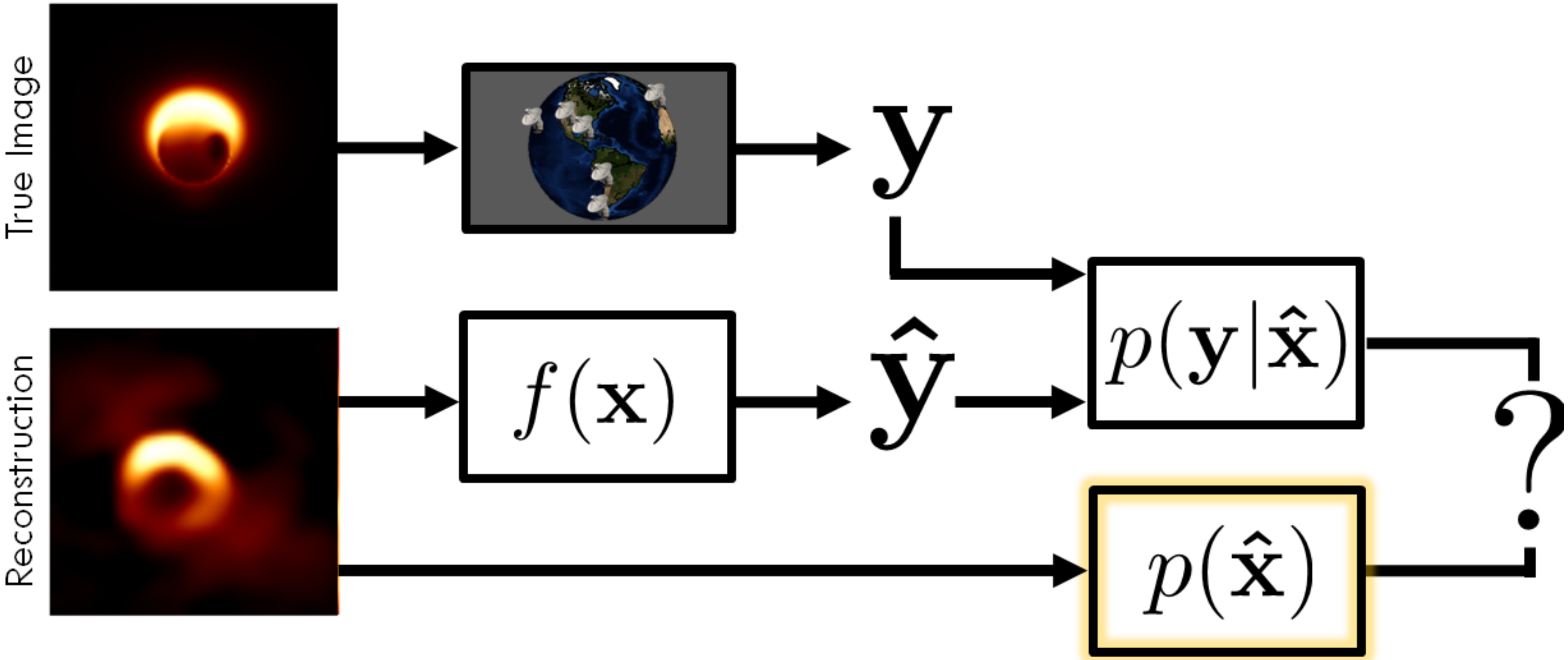
8.3.3 Imaging the Invisiable



8.3. Regularization

8.3.3 Imaging the Invisiable

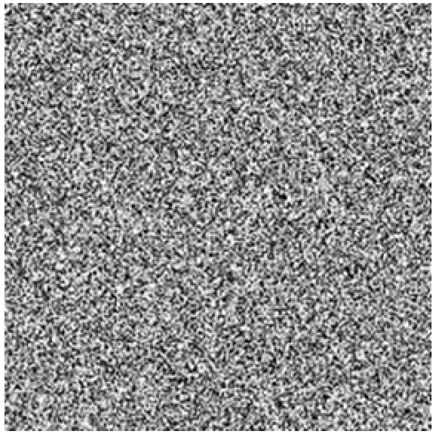
Bayesian Model Inversion



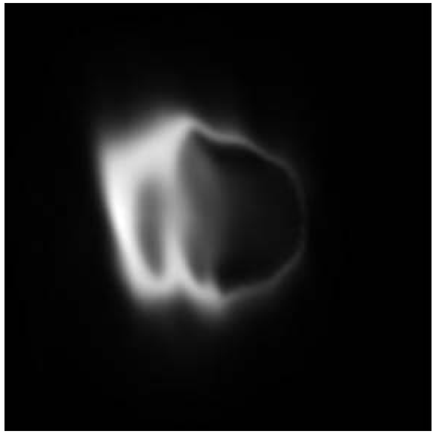
8.3. Regularization

8.3.3 Imaging the Invisiable

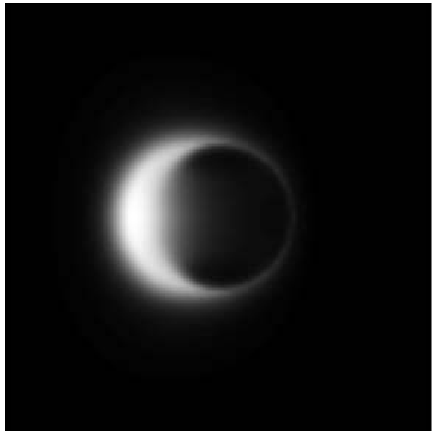
$$\overbrace{\hat{\mathbf{x}}_{\text{MAP}}}^{\text{Best Image}} = \text{argmax}_{\mathbf{x}} \left[\log \underbrace{p(\mathbf{y}|\mathbf{x})}_{\text{Likelihood}} + \log \underbrace{p(\mathbf{x})}_{\text{Prior}}\right]$$



Unlikely



More likely



Very likely

8.3. Regularization

8.3.3 Imaging the Invisiable

Bayesian Model Inversion

Systematic Gain Error

Thermal Noise

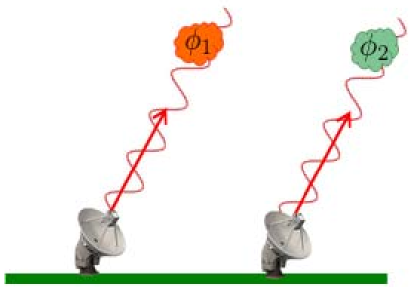
Quantization Error

Atmospheric Error

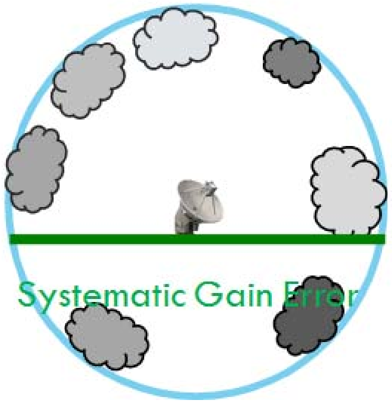
The Astronomical Source's Dynamics

Scattering

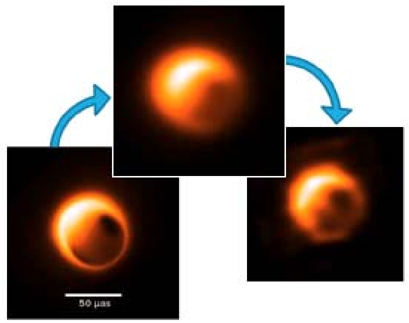
$$\hat{\mathbf{x}}_{\mathrm{MAP}} = \operatorname{argmax}_{\mathbf{x}} \left[\log \boxed{p(\mathbf{y}|\mathbf{x})} + \log p(\mathbf{x})\right]$$



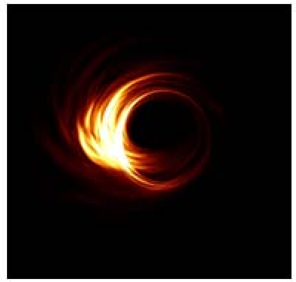
Atmospheric Error



Systematic Gain Error



Scattering

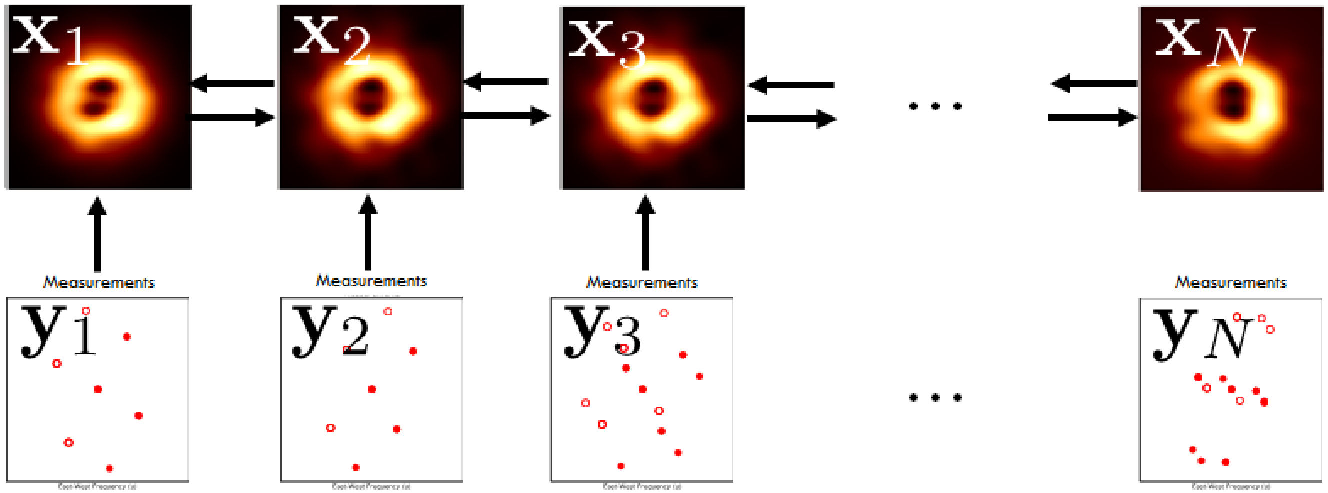


Source Dynamics

8.3. Regularization

8.3.3 Imaging the Invisiable

$$p(\mathbf{X}, \mathbf{Y}) \propto \prod_{t=1}^{N} \varphi_{\mathbf{y}_t|\mathbf{x}_t} \prod_{t=1}^{N} \varphi_{\mathbf{x}_t} \prod_{t=2}^{N} \varphi_{\mathbf{x}_t|\mathbf{x}_{t-1}}$$

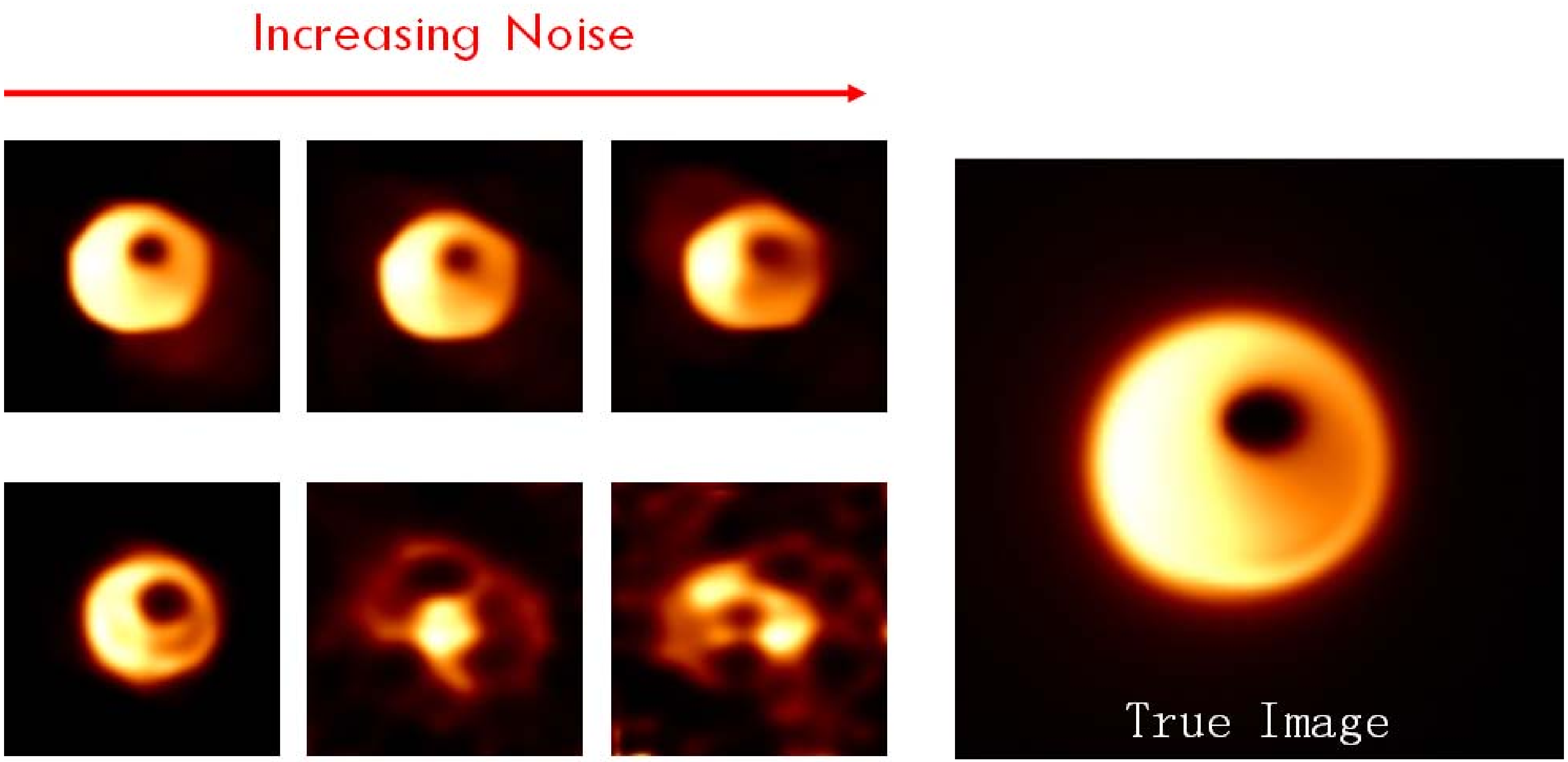


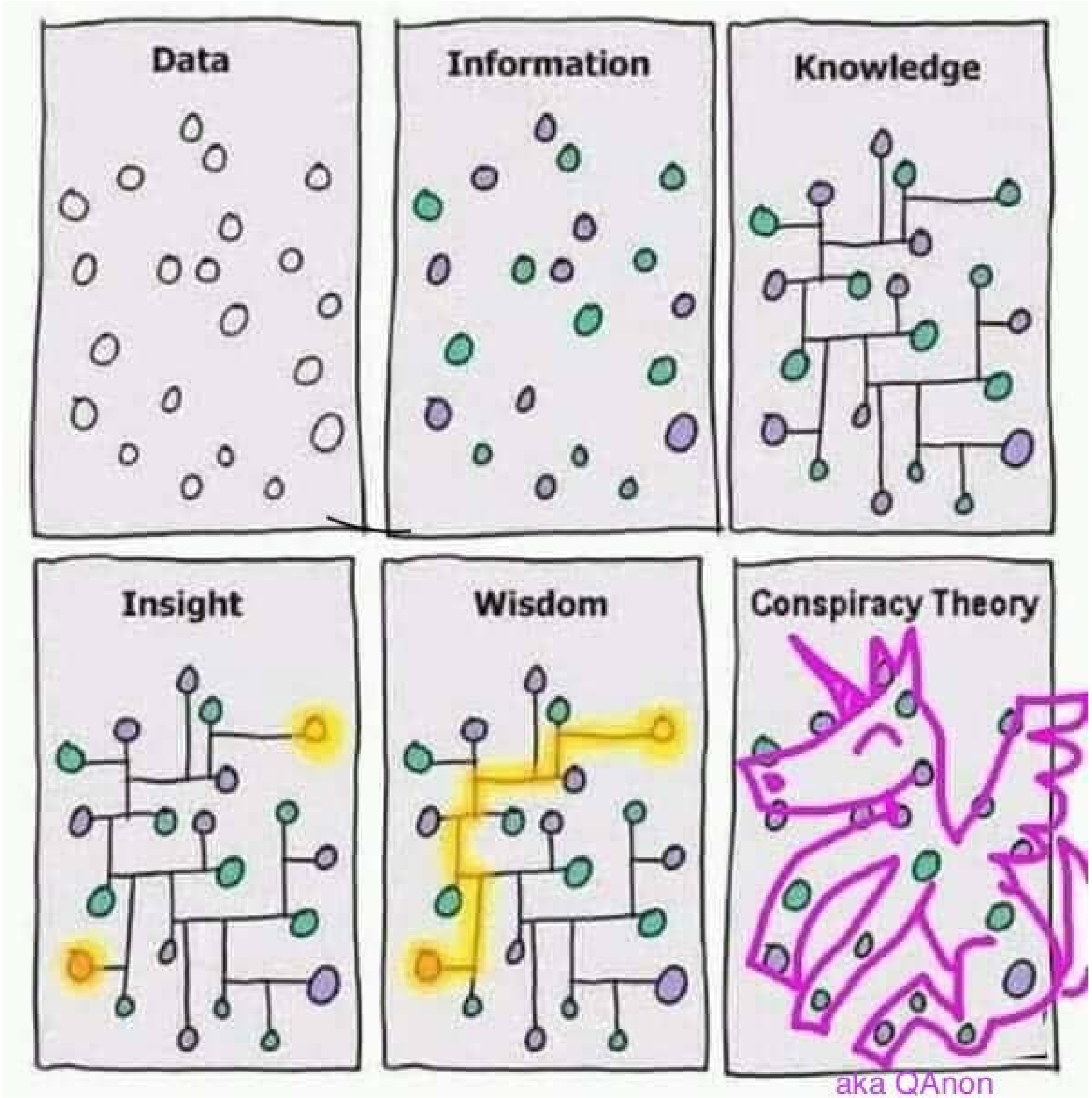
Each video frame should look similar to its adjacent video frames

8.3. Regularization

8.3.3 Imaging the Invisiable

一言以蔽之，“胸有成竹”




Data
Information
Knowledge
Insight
Wisdom
Conspiracy Theory
aka QAnon

8.4. References

[1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B Statistical Methodology*, vol. 58, no. 1, pp. 267-288, 1996.

[2] S. S. Chen, D. L. Donoho, M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129-159, 2001.

[3] D. L. Donoho, X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Transactions on Information Theory*, vol. 47, no. 7, pp. 2845-2862, 2001.

[4] D. L. Donoho, M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via l_1 minimization," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197-2202, 2003.

[5] E. Candes, M. Rudelson, T. Tao, R. Vershynin, "Error correction via linear programming," *Proceedings of IEEE Symposium on Foundations of Computer Science*, pp. 295-308, 2004.

[6] E. Candés, T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203-4215, 2004.

[7] E. Candés, T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406-5425, 2006.

[8] E. Candés, J. Romberg, T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207-1223, 2006.

[9] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289-1306, 2006.

[10] B. A. Cipra, "l_1-Magic," *SIAM News*, vol. 39, no. 9, 2006.

[11] E. Candés, T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Annals of Statistics*, vol. 35, no. 6, pp. 2313-2351, 2007.

[12] E. J. Candes, T. Tao, "The power of convex relaxation: Near-Optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053-2080, 2010.

[13] S. Shalev-Shwartz, N. Srebro, T. Zhang, "Trading accuracy for sparsity in optimization problems with sparsity constraints," *SIAM Journal of Optimization*, vol. 20, no. 6, pp. 2807-2832, 2010.

[14] P. Tseng, "Approximation accuracy, gradient methods, and error bound for structured convex optimization," *Mathematical Programming*, vol. 125, no. 2, pp. 263-295, 2010.

[15] A. Juditsky, A. Nemirovski, "On verifiable sufficient conditions for sparse signal recovery via l_1 minimization," *Mathematical Programming*, vol. 127, no. 1, pp. 57-88, 2011.

[16] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, Springer, 2010.

[17] Z. Luo, W. Yu, "An introduction to convex optimization for communications and signal processing," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1426-1438, 2006.

[18] A. M. Bruckstein, D. L. Donoho, M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Review*, vol. 51, no. 1, pp. 34-81, 2009.

[19] J. Mattingely, S. Boyd, "Real-Time convex optimization in signal processing," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 50-61, 2010.

[20] H. Jiang, X. Li, M. Zibulevsky, M. Elad, "L1-L2 optimization in signal and image processing," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 76-88, 2010.

[21] H. Jiang, X. Li, "Parameter estimation of statistical models using convex optimization," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 115-127, 2010.

[22] M. Lustig, D. L. Donoho, J. M. Santos, J. M. Pauly, "Compressed sensing MRI," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 72-82, 2008.

[23] E. J. Candès, Y. Plan, "A probabilistic and RIPless theory of compressed sensing," *IEEE Transactions on Information Theory*, vol. 57, no. 11, pp. 7235-7254, 2011.

[24] A. N. Tychonoff, "Об устойчивости обратных задач [On the stability of inverse problems]," *Doklady Akademii Nauk SSSR*, vol. 39, no. 5, pp. 195-198, 1943.

[25] A. E. Hoerl, "Application of ridge analysis to regression problems," *Chemical Engineering Progress*, vol. 58, pp. 54-59, 1962.

[26] A. E. Hoerl, R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 42, no. 1, pp. 80-86, 1970.

[27] A. J. Smola, B. Schölkopf, K.-R. Müller, "The connection between regularization operators and support vector kernels," *Neural Networks*, vol. 11, no. 4, pp. 637-649, 1998.

[28] P. J. Bickel, B. Li, A. B. Tsybakov, S. A. van de Geer, B. Yu, T. Valdés, C. Rivero, J. Fan, A. van der Vaart, "Regularization in statistics," *Test*, vol. 15, no. 2, pp. 271-344, 2006.

[29] M. E. Tipping, "Bayesian inference: An introduction to principles and practice in machine learning," *Advanced Lectures on Machine Learning*, O. Bousquet, U. von Luxburg, G. Ratsch, eds., pp. 41-62, Springer, 2002. http://www.miketipping.com/papers/met-mlbayes.pdf

[30] K. L. Bouman, *Extreme Imaging via Physical Model Inversion: Seeing Around Corners and Imaging Black Holes*, Ph.D. Dissertation, MIT, 2017.

[31] D. Castelvecchi, "Black hole imaged for first time," *Nature*, vol. 568, pp. 284-285, 2019.

[32] https://baijiahao.baidu.com/s?id=1630434972238556896