Limits to Nonlinear Inversion

Klaus Mosegaard

Univ. of Copenhagen

September 2008

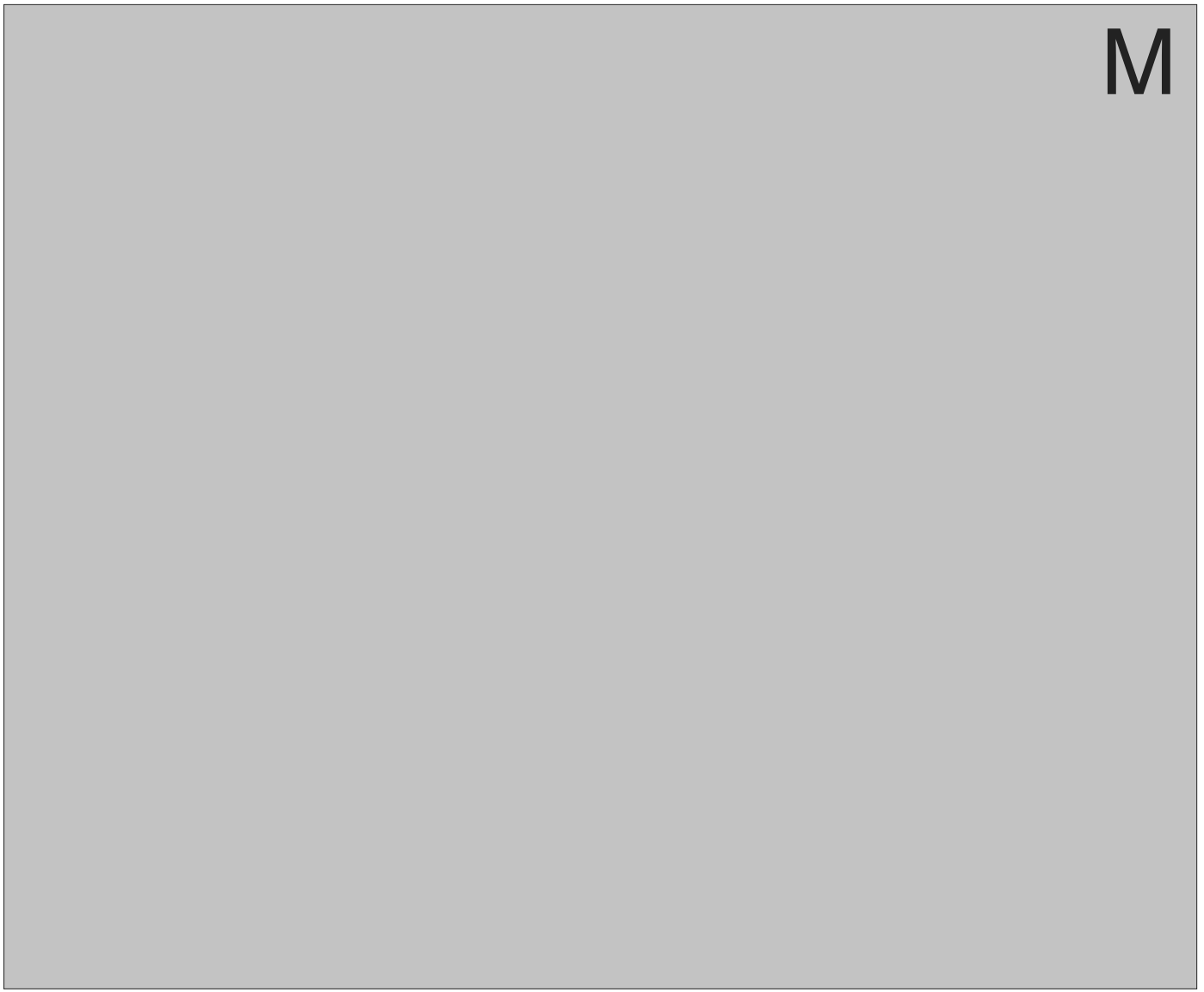
Outline (and basic theses to be substantiated)

1. The most difficult task: **To find a solution!**.
2. Once the solutions are found, **evaluation of uncertainties** is usually relatively easy.
3. If the inversion algorithm has not converged properly to the solution(s), this is the **most significant source of uncertainty**!
4. The futility of **blind inversion** - the use of general purpose algorithms.
5. Inversion **algorithms built for the specific problem** perform better!

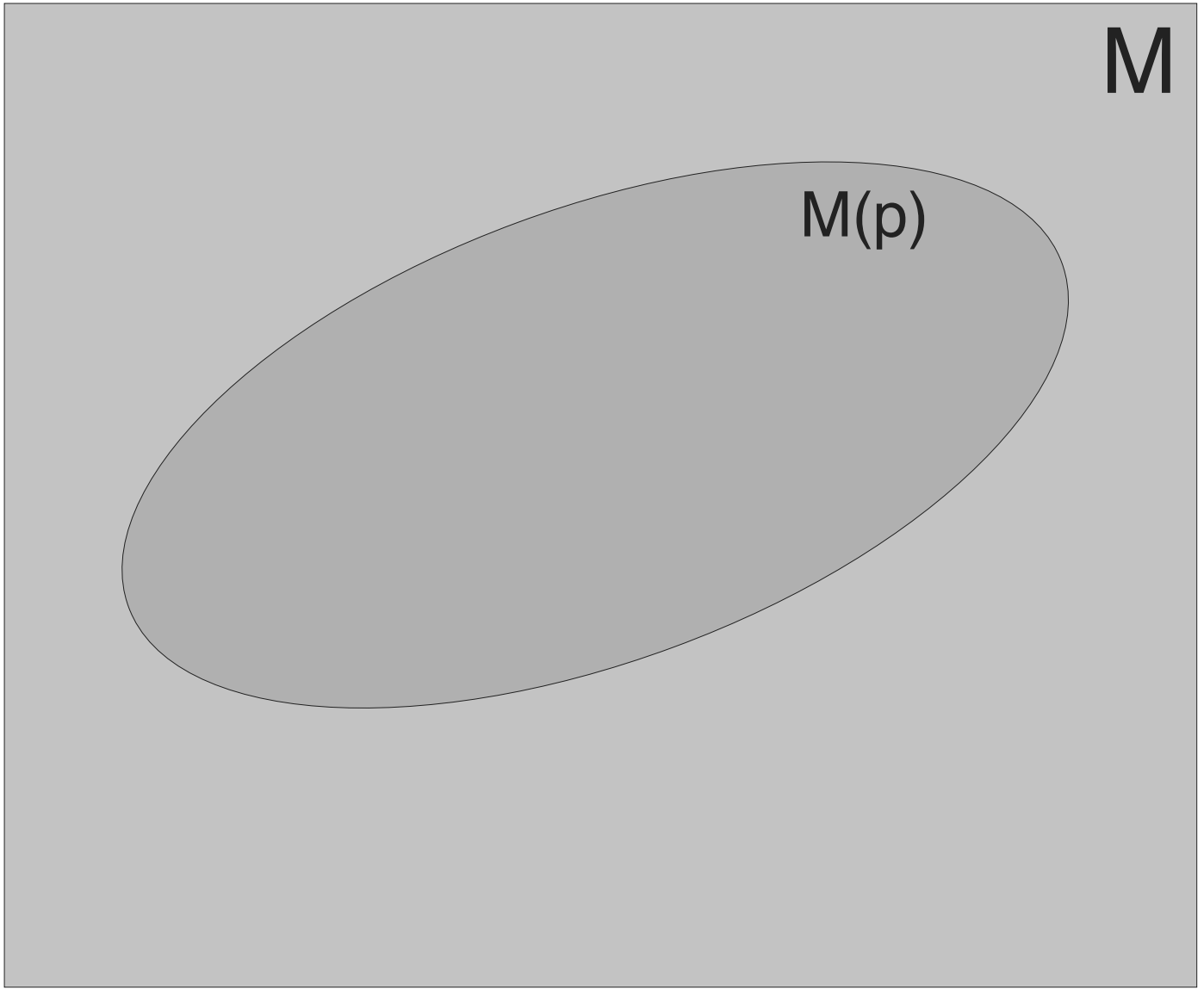
The most difficult problem:

To find a solution!

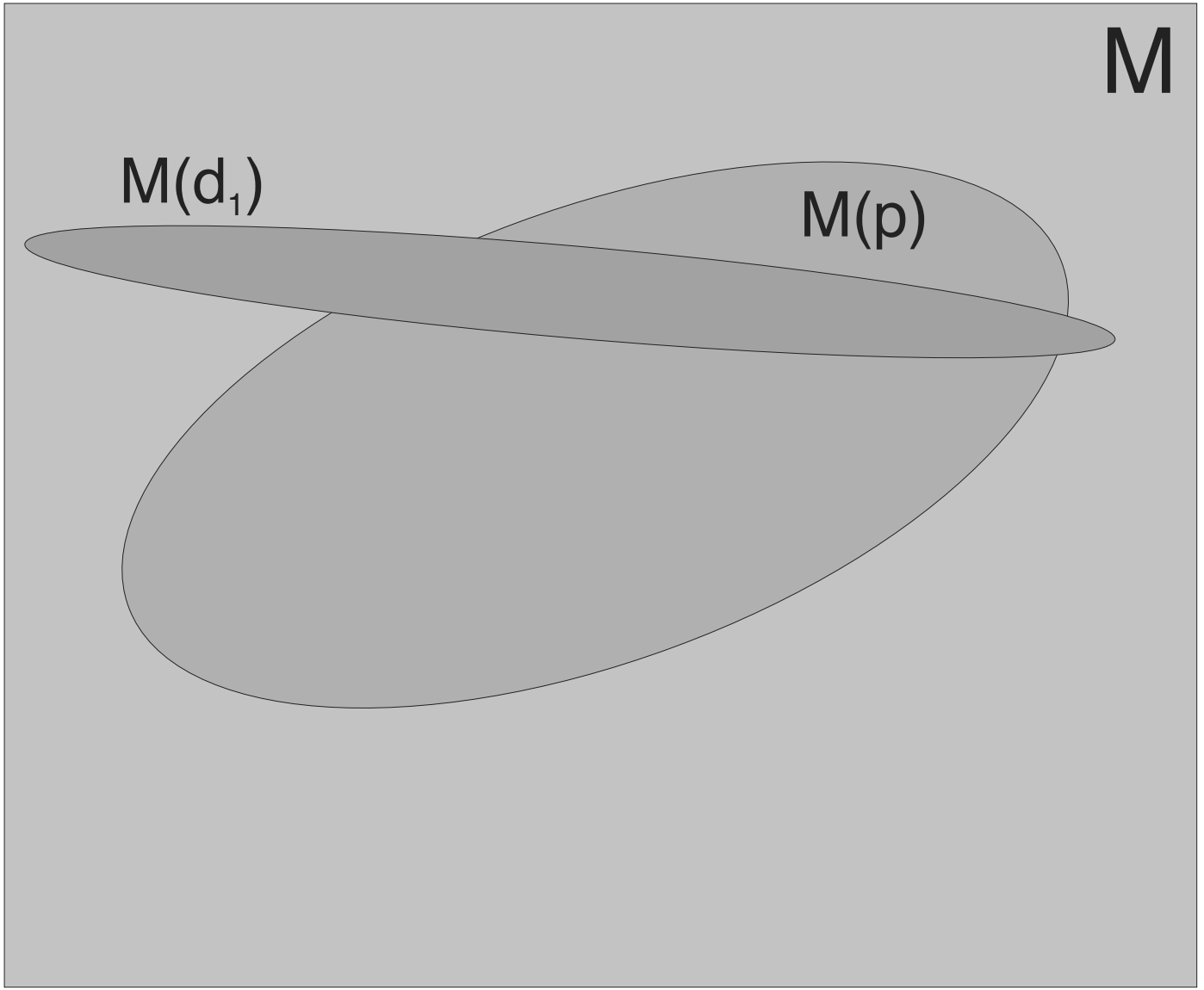
The logic of Data Analysis



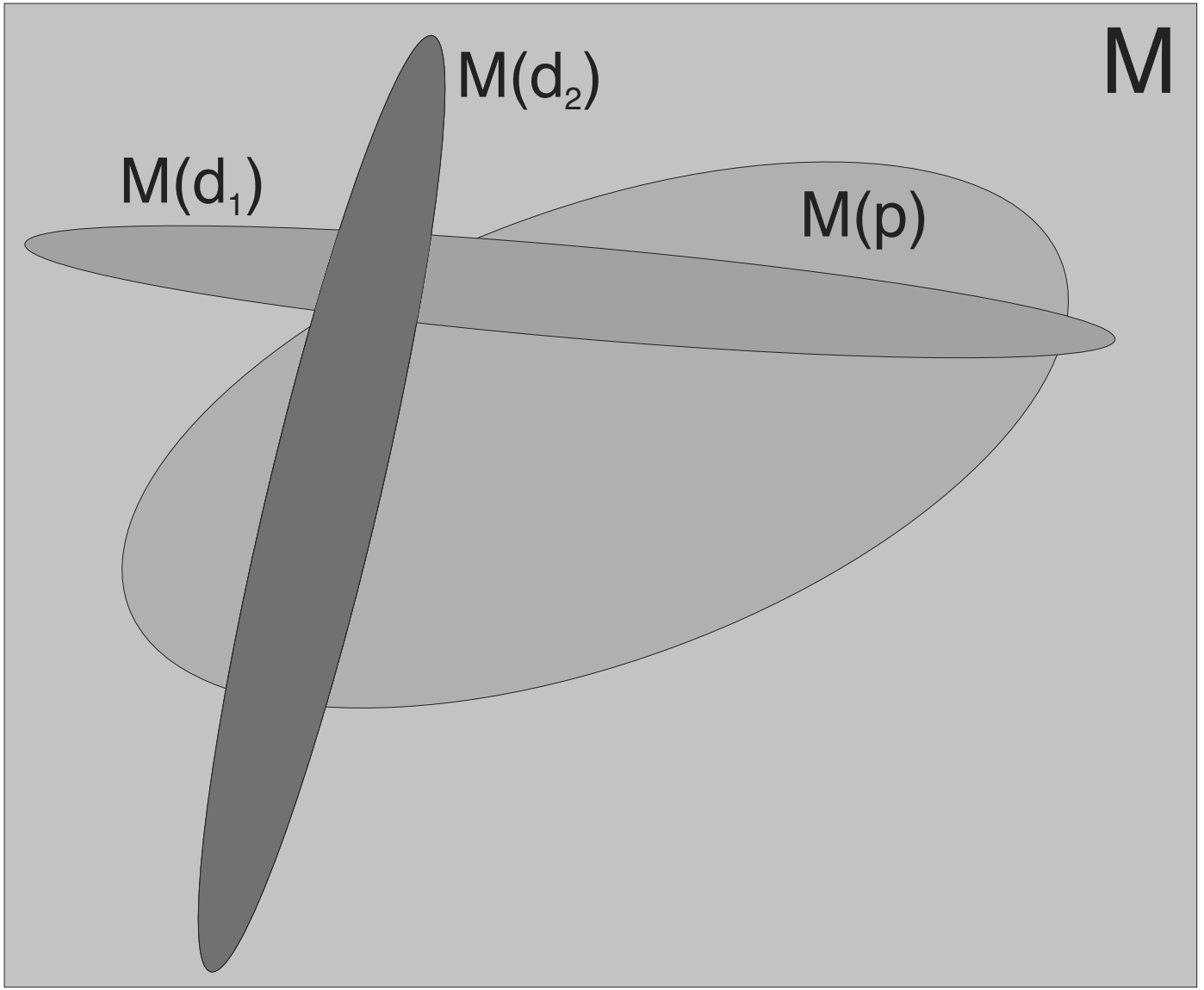
The logic of Data Analysis



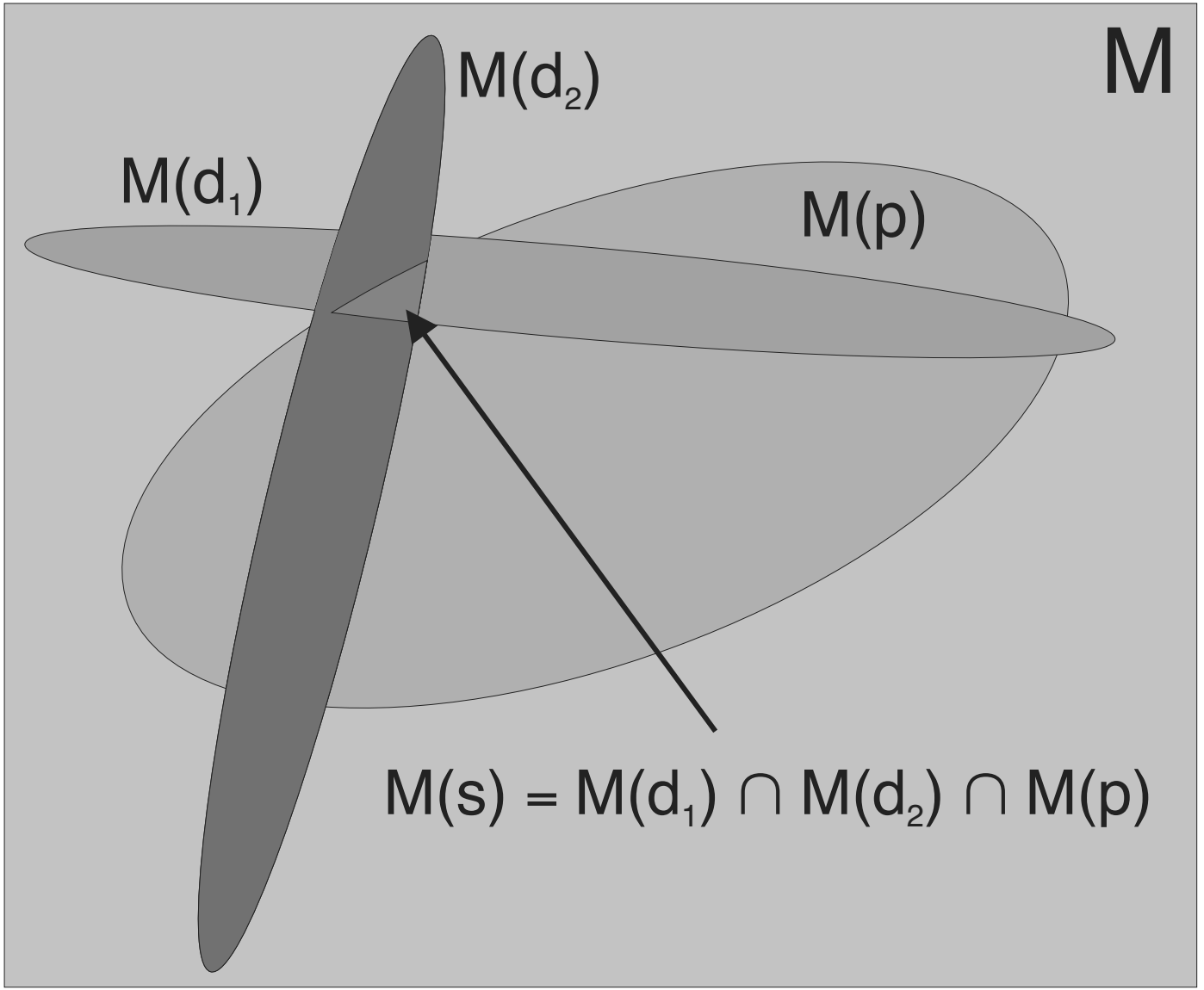
The logic of Data Analysis



The logic of Data Analysis



The logic of Data Analysis



The Bayesian view

Define indicator functions:

$$L_j(\mathbf{m}) = \begin{cases} 1 \text{ if } \mathbf{m} \in M(d_j) \\ 0 \text{ otherwise} \end{cases}$$

$$\rho(\mathbf{m}) = \begin{cases} 1 \text{ if } \mathbf{m} \in M(p) \\ 0 \text{ otherwise} \end{cases}$$

$$\sigma(\mathbf{m}) = \begin{cases} 1 \text{ if } \mathbf{m} \text{ is a solution} \\ 0 \text{ otherwise} \end{cases}$$

then

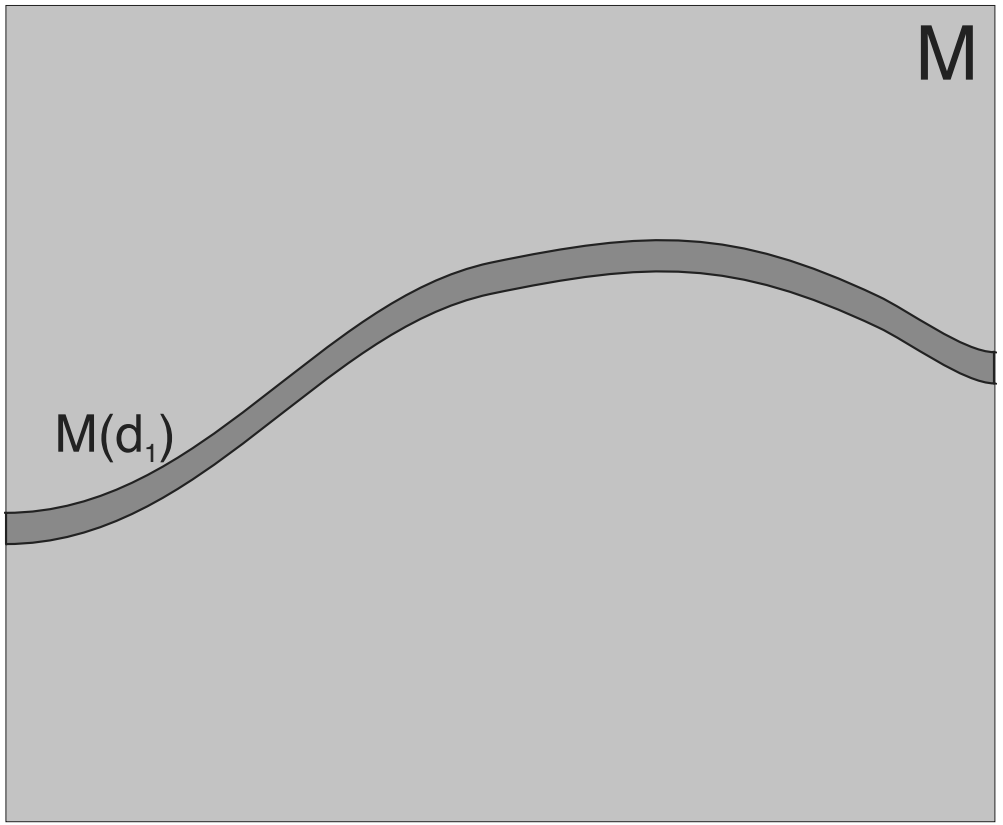
$$\sigma(\mathbf{m}) = \underbrace{L_1(\mathbf{m}) \dots L_N(\mathbf{m})}_{L(\mathbf{m}|\mathbf{d})} \ \rho(\mathbf{m})$$

"Softening" the indicator functions to probability densities leaves us with Bayes' Rule.

The Deterministic view

Models consistent with **one datum** usually reside in a "narrow neighbourhood" of a manifold with dimension

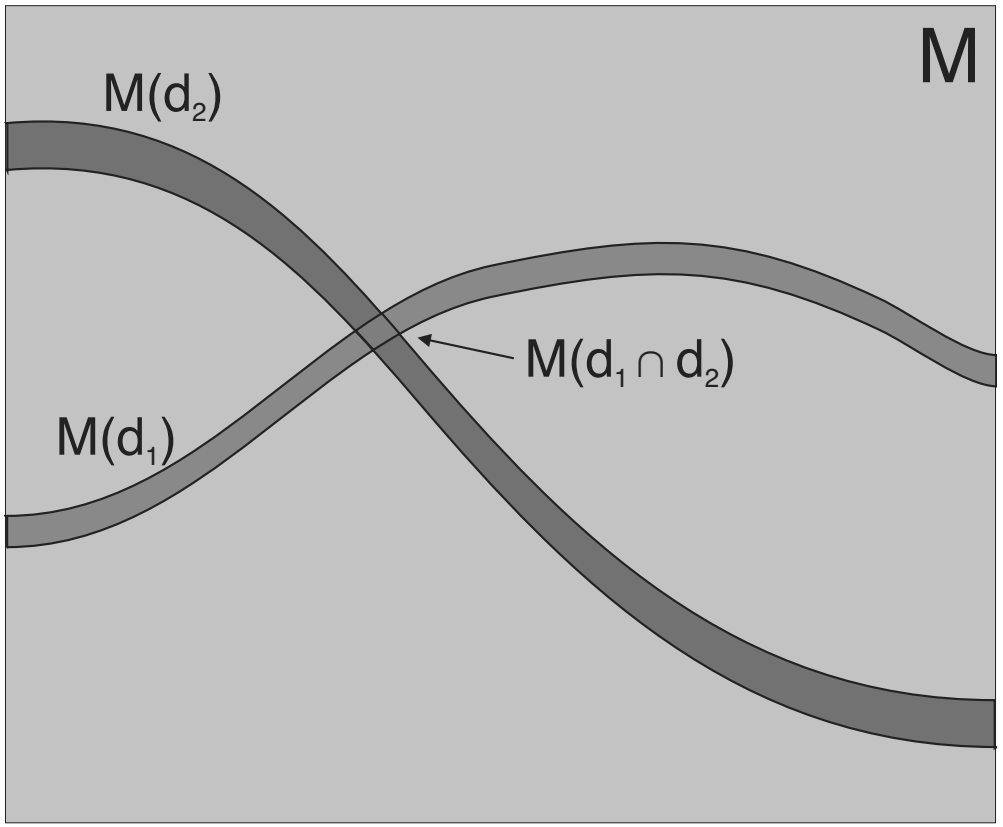
$$\mathrm{Dim}(M) - 1$$



The Deterministic view

Models consistent with $\mathbf{N}$ **independent data** usually reside in a "narrow neighbourhood" of a manifold with dimension

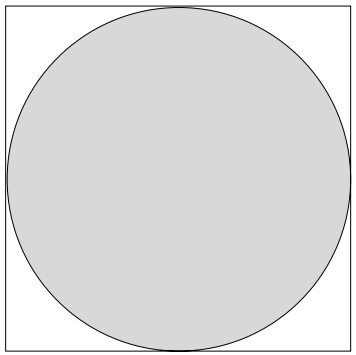
$$\mathrm{Dim}(M) - N$$

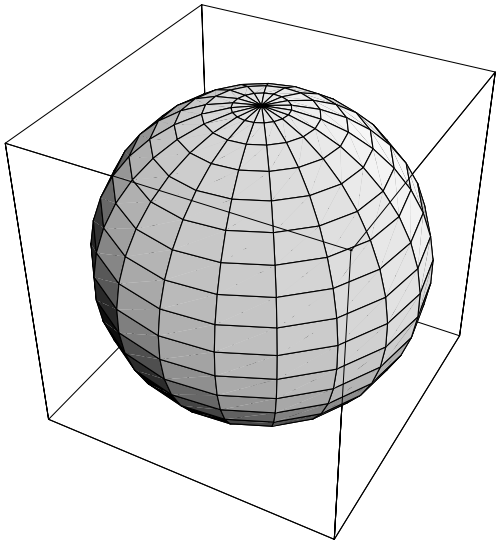


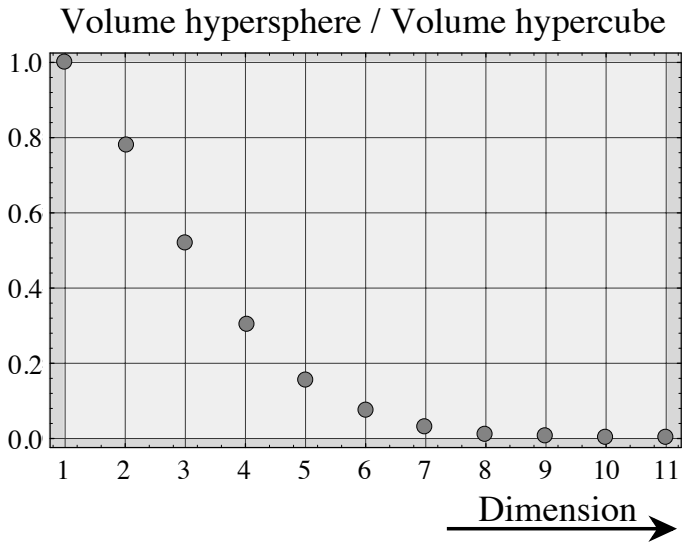
The "Curse of Dimensionality"

The volume of the solution space decreases at least exponentially with the number of independent data

2R	πR^2	$\frac{4}{3}\pi R^3$	(...)	$\frac{\pi^n}{n!}R^{2n}$	$\frac{2^{n+1}\,\pi^n}{(2n+1)!!}R^{2n+1}$
2R	$(2R)^2$	$(2R)^3$	(...)	$(2R)^{2n}$	$(2R)^{2n+1}$







Preliminary observations

Let

$\text{Dim}(M)$: Dimension of model parameter space
$\text{Dim}(D)$: Dimension of data space
$\text{Dim}(P)$: Number of independent a priori constraints

Observation 1 Given the path to a point in the solution space, the **search time along the path** is only weakly dependent on $\text{Dim}(M)$, $\text{Dim}(D)$ and $\text{Dim}(P)$.

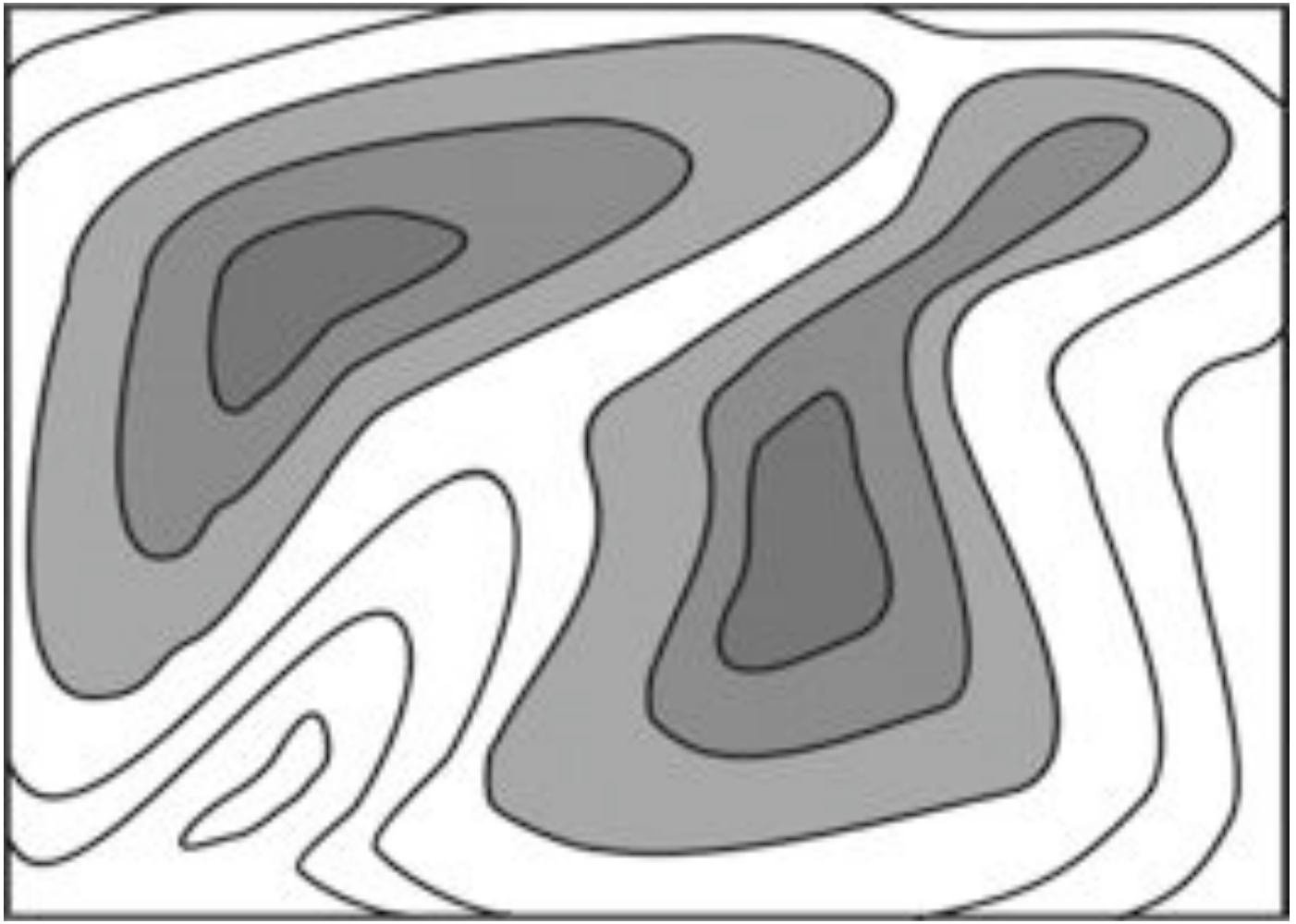
Observation 2
Given no information about the solution space, the **random search time** increases at least exponentially with

$$\text{Dim}(M) + \text{Dim}(D) + \text{Dim}(P) \tag{1}$$

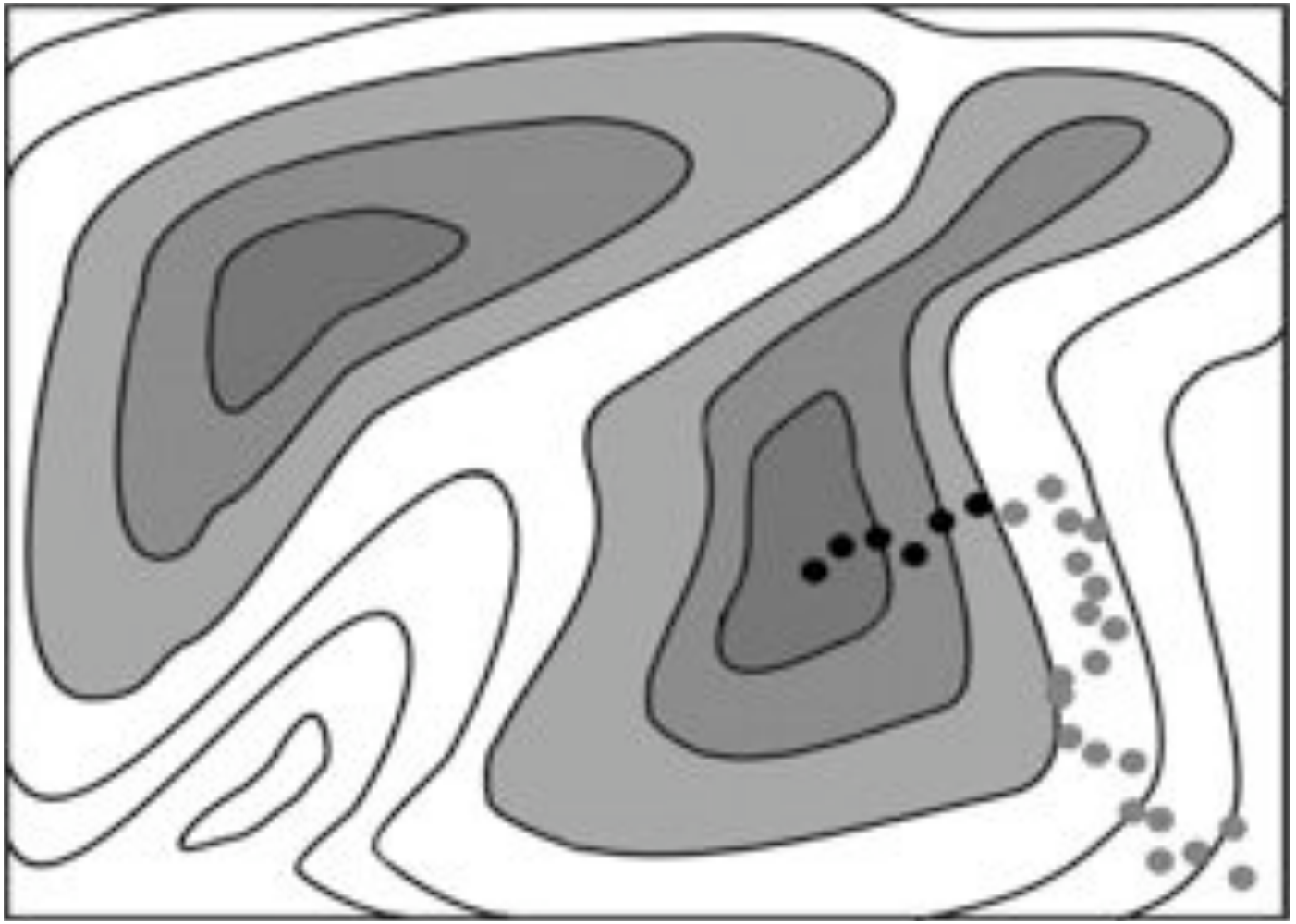
when $\text{Dim}(M) \geq \text{Dim}(D)$

Once the solutions are found, evaluation of uncertainties, is usually relatively easy!

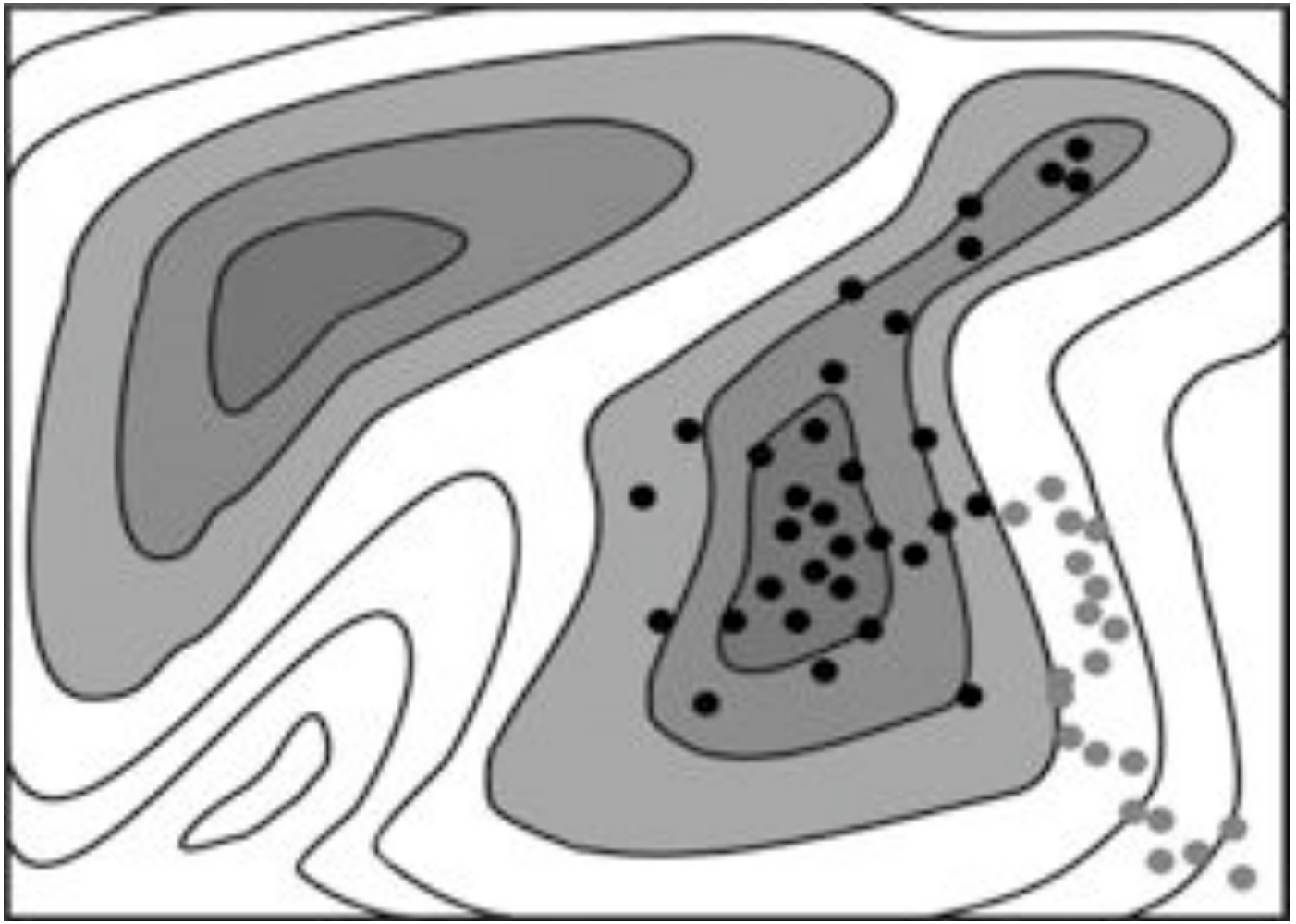
Search and sampling



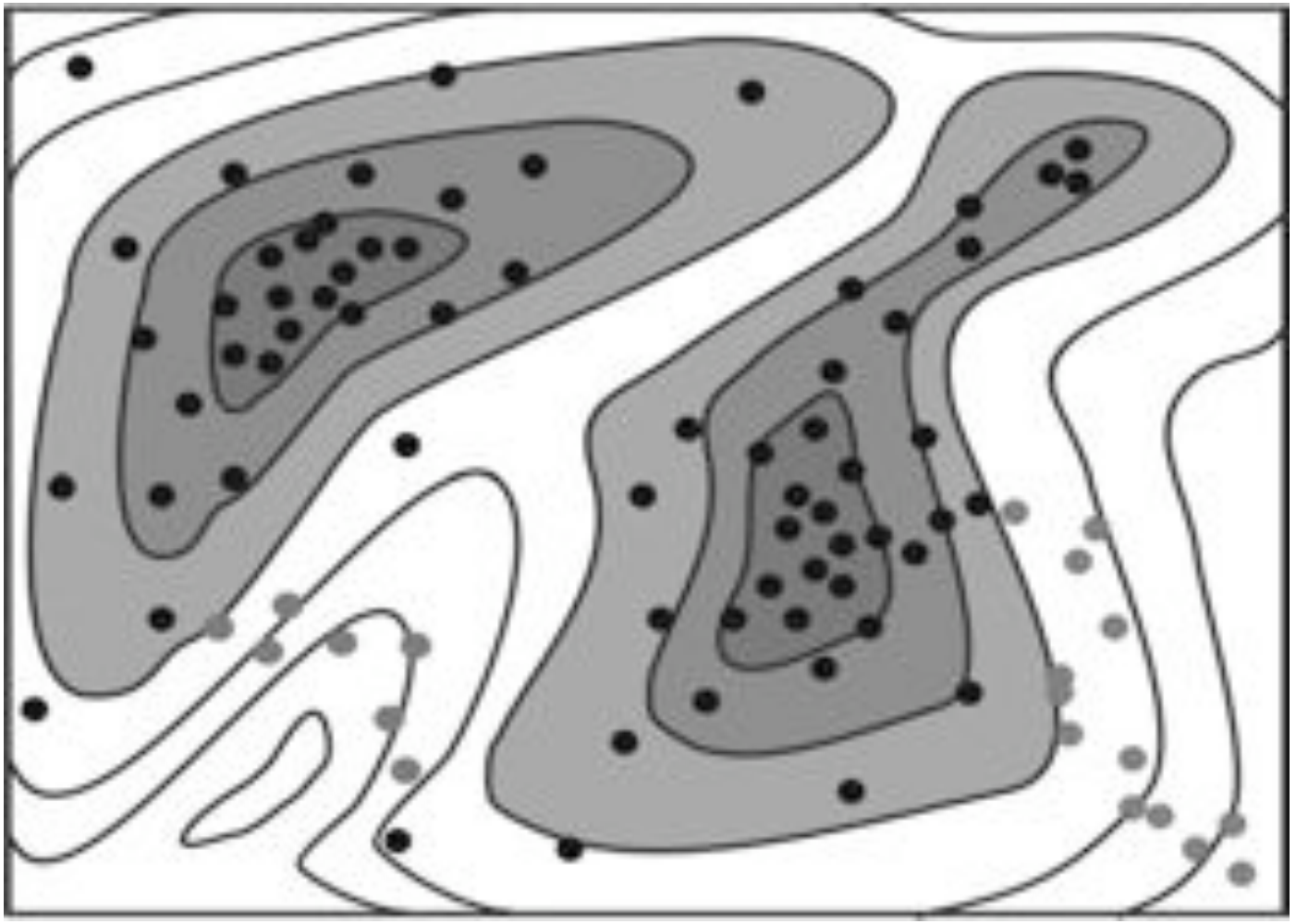
Search and sampling



Search and sampling

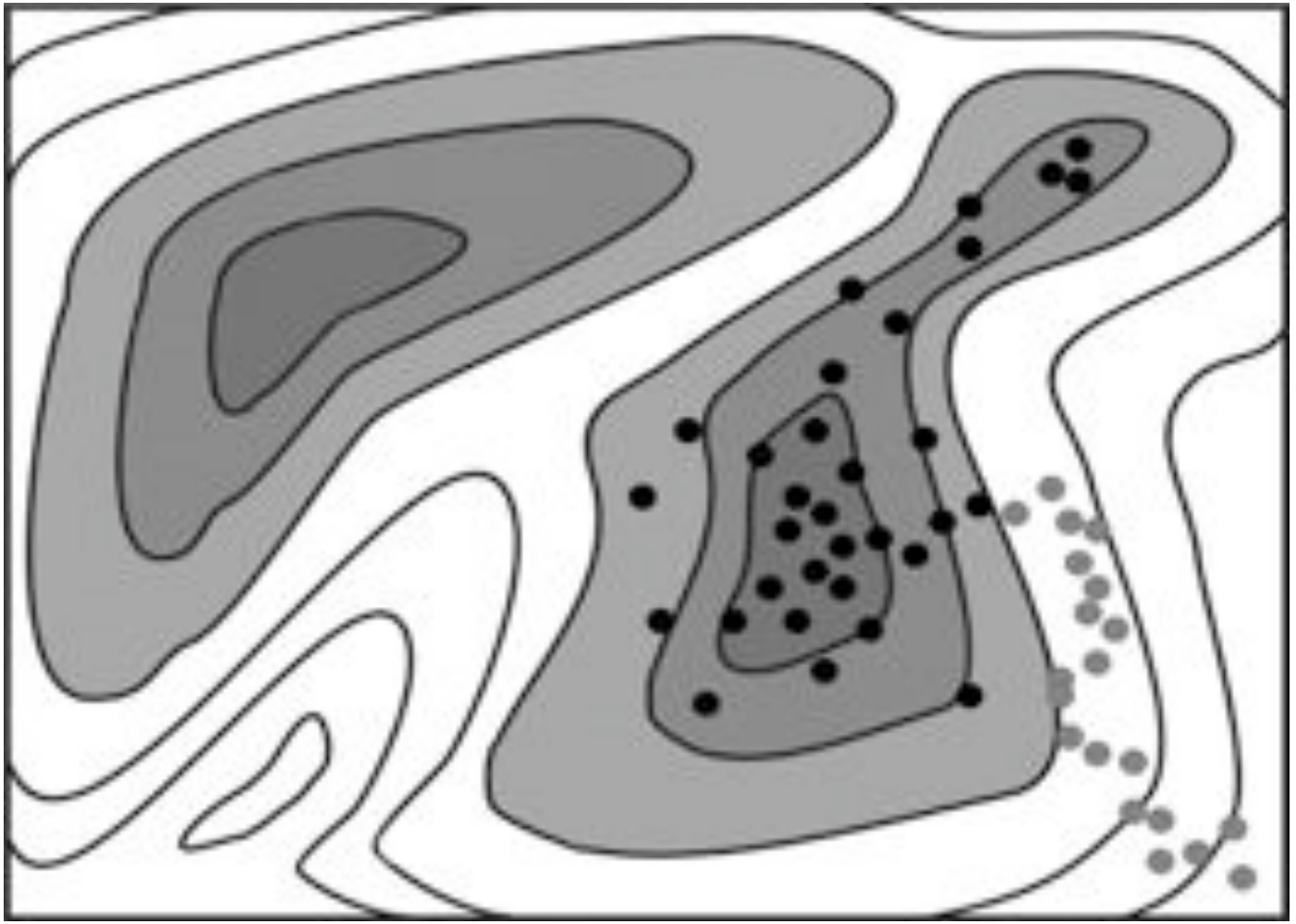


Search and sampling



If the inversion algorithm has not converged properly to the solution(s), this is the most significant source of uncertainty!

Incomplete convergence



The futility of blind inversion

The question

Which one of the following general purpose algorithms is the most efficient?

- Simulated Annealing,
- Metropolis Algorithm,
- Random Search,
- Rejection Sampling,
- Genetic Algorithm,
- Taboo Search,
- Neighbourhood Algorithm,
- ...

?

A different viewpoint:

Double-discrete Analysis of Inverse Problems

Double-discrete data analysis

Here, we shall assume that model parameters are **doubly discrete**:

- There is a **finite number of model parameters** (this is the usual assumption in parameter estimation)
- Model parameters can only take a **finite number of parameter values**!

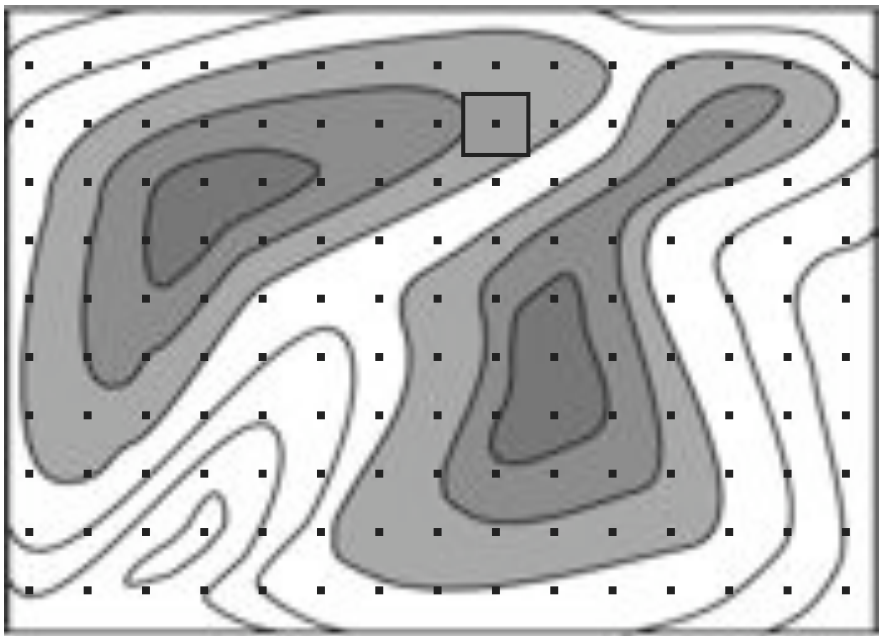
Figure: Original image, image with few pixels, and image with few color levels

How fine a discretization is needed for an inverse problem?

- The misfit function $f(\mathbf{m})$ usually inherits continuity from $\mathbf{d} = g(\mathbf{m})$, e.g.,

$$f(\mathbf{m}) = \frac{\|\mathbf{d} - g(\mathbf{m})\|^2}{2\sigma^2}$$

- Now we can define **a grid of points** representing **small regions** $\Delta m_1 \Delta m_2 \ldots$ of **almost constant** $f(\mathbf{m})$.



How fine a discretization of parameters values is needed?

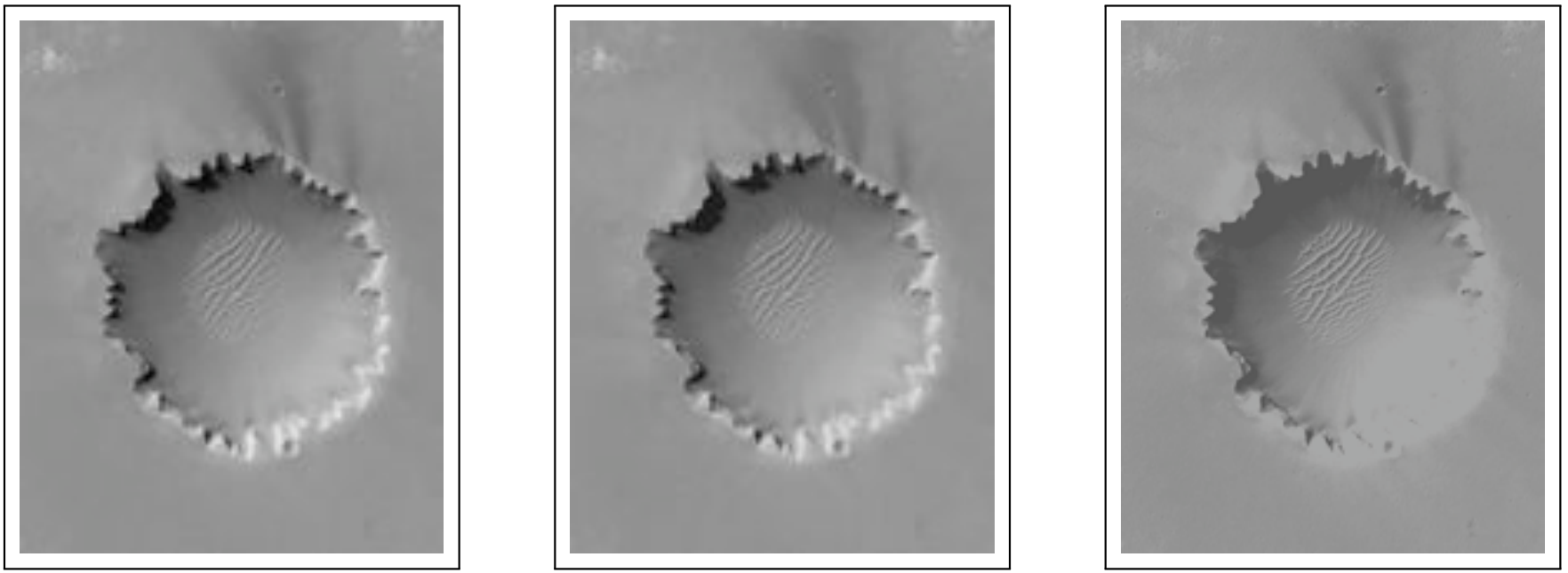
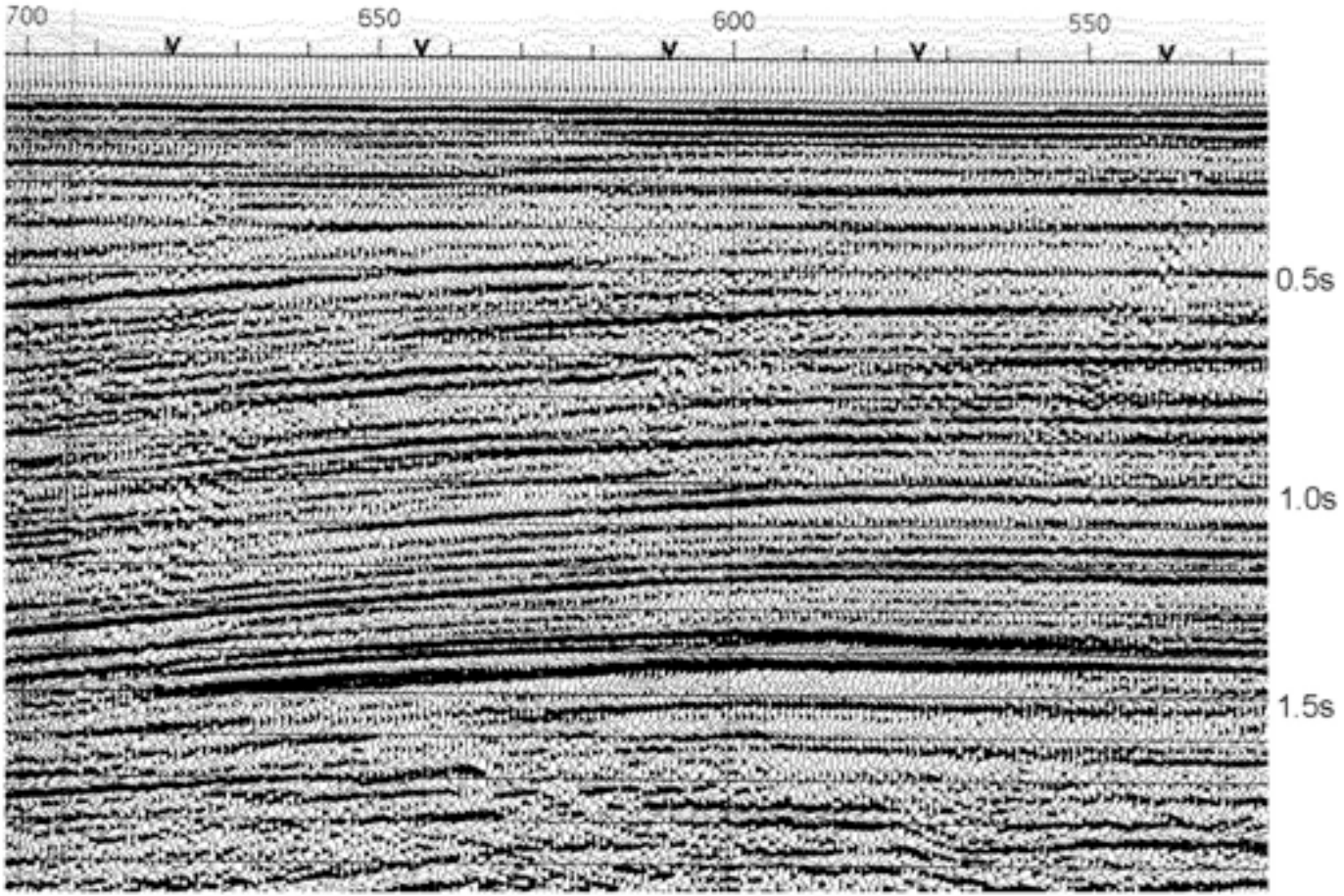


Figure: The Victoria Crater in 256 colors, 16 colors, and 4 colors.

Example: Seismic reflection data



- $\Delta m_i < 2\sigma^2\epsilon/\|\mathbf{w}\|^2$, where σ is the standard deviation of the noise, ϵ is the desired fractional change in misfit over Δm_i, and $\mathbf{w}$ is the seismic wavelet.

The discrete counterpart to "The Curse of Dimensionality"

- The inverse problem:

$$
\begin{aligned}
d_1 &= g_1(m_1, m_2 \ldots, m_M) \\
d_2 &= g_2(m_1, m_2 \ldots, m_M) \\
&\vdots \\
d_K &= g_K(m_1, m_2 \ldots, m_M)
\end{aligned}
$$

- Here, we can freely chose one out of N values for $M - K$ model parameters. This can be done in N^{M-K} ways.
- After this we have K equation with K unknowns left, and they may have a solution in one, several or all of the above N^{M-K} cases.

Proposition

The curse of combinatorics. *K data reduce the solution space by a factor $\leq N^{-K}$*

A Double-discrete Analysis of the Performance of Inversion Algorithms

The typical scenario for nonlinear inversion

- In the relations

$$d_i = g_i(\mathbf{m}).$$

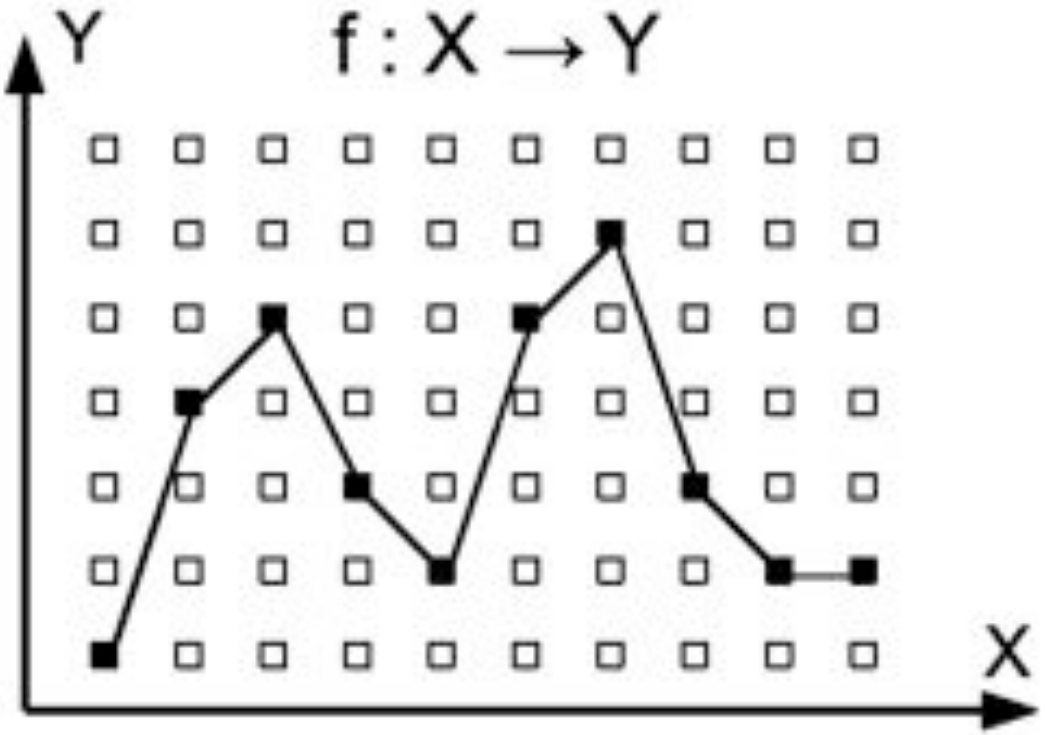
  we have **no closed-form** mathematical expression for $g_i(\mathbf{m})$.
- We only have a programme that is able to **evaluate** $g_i(\mathbf{m})$ **for given values** of the parameters in $\mathbf{m}$.

In short:

We are performing a **blind search** for the solution.

Notation 1

- Two *finite sets* X and Y,
- The set $\mathcal{F}_X$ of all fit functions/probability distributions $f : X \to Y$.

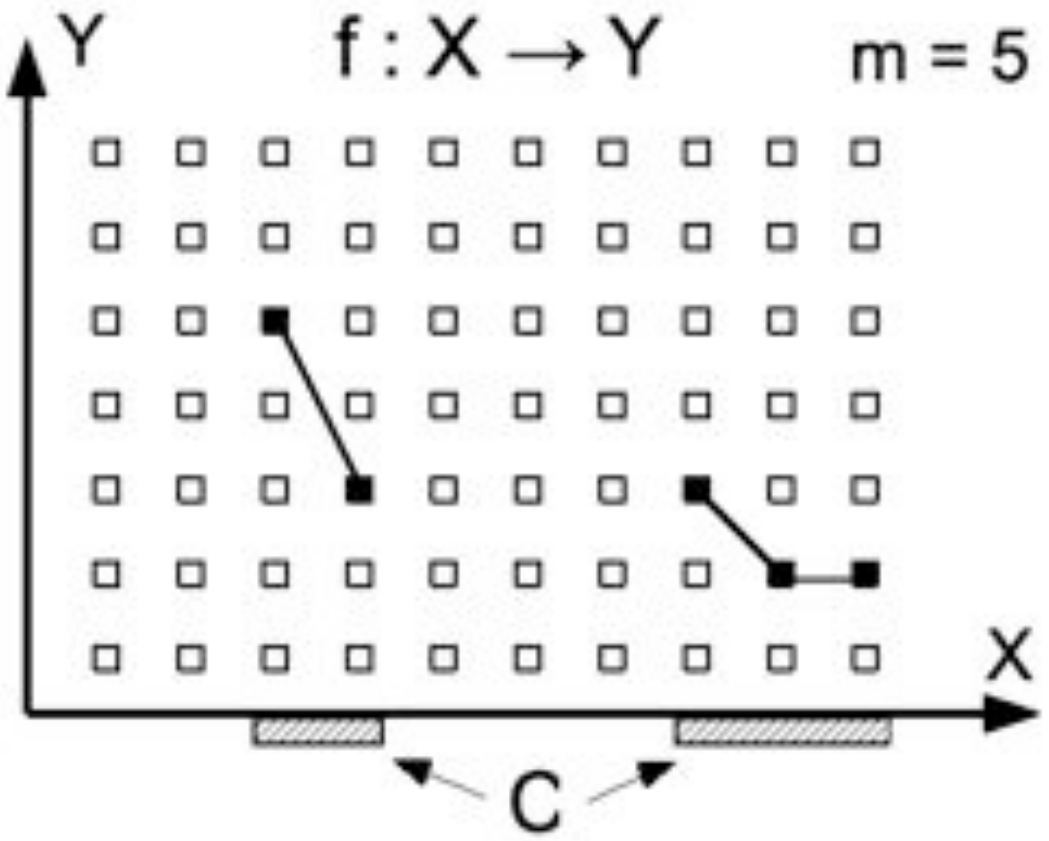


Notation 2

- A sample of size $m < |X|$:

$$\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}.$$

- The set $\mathcal{F}_{X|C}$ of all fit functions/probability distributions defined on X, but with fixed values in C.



Proposition

The total number of functions intersecting the m samples is

$$|\mathcal{F}_{X|C}| = |Y|^{|X|-m}. \tag{2}$$

This number is independent of the location of the sample points.

The probability that an algorithm a sees the values $y_1, \ldots, y_m$ in the first m steps is then

$$P(y_1, \ldots, y_m | f, m, a) = \frac{|Y|^{|X|-m}}{|Y|^{|X|}} = |Y|^{-m} \tag{3}$$

This number is independent of the algorithm.

The No-Free-Lunch Theorem adapted to inversion

Theorem

NFL (Wolpert and Macready, 1995) *For $f \in \mathcal{F}_X$ and any pair of algorithms a_1 and a_2,*

$$P(y_1, \ldots, y_m | f, m, a_1) = P(y_1, \ldots, y_m | f, m, a_2) \tag{4}$$

where $P(\cdot|\cdot)$ denotes conditional probability.

Corollary

(NFL for optimization) *When all fit functions are equally probable (blind inversion), the distribution of any performance measure $\Phi(y_1, \ldots, y_m)$ for inversion is exactly the same for all inversion algorithms.*

A simple performance measure for inversion could be $\Phi(y_1, \ldots, y_m) = \max\{y_1, \ldots, y_m\}$ which must be large for good performance.

Critique of the NFL theorem

Postulate

Our **fit functions** *(misfit functions or probability densities) belong to a* **narrow family** *of functions (e.g., smooth functions), and some algorithms work better than others on such families!*

So, the situation is different from the NFL-scenario:

We have a narrow set of functions (albeit unknown to the algorithm).

We can, however, extend the NFL Theorem to the following

Theorem

The average performance over all fit function families is exactly the same for all inversion algorithms.

Conclusion

The efficiency of all **blind** inversion schemes:

- Simulated Annealing,
- Metropolis Algorithm,
- Genetic Algorithm,
- Taboo Search,
- Neighbourhood Algorithm,
- . . . ,

when averaged over alle conceivable inverse problems, are exactly the same.

A final corollary

Corollary

Only an algorithm **adapted to the specific problem** *has a chance of performing better than a random search.*

In fact, the following theorem can be demonstrated:

Theorem

A step length of $2n + 1$*, where* n *is the correlation distance of the fit function, is optimally reducing the set of possible solutions.*