Statistical Measures of Uncertainty in Inverse Problems

Workshop on Uncertainty in Inverse Problems

Institute for Mathematics and Its Applications

Minneapolis, MN 19-26 April 2002

P.B. Stark
Department of Statistics
University of California
Berkeley, CA 94720-3860
www.stat.berkeley.edu/~stark

Abstract

Inverse problems can be viewed as special cases of statistical estimation problems. From that perspective, one can study inverse problems using standard statistical measures of uncertainty, such as bias, variance, mean squared error and other measures of risk, confidence sets, and so on. It is useful to distinguish between the intrinsic uncertainty of an inverse problem and the uncertainty of applying any particular technique for "solving" the inverse problem. The intrinsic uncertainty depends crucially on the prior constraints on the unknown (including prior probability distributions in the case of Bayesian analyses), on the forward operator, on the statistics of the observational errors, and on the nature of the properties of the unknown one wishes to estimate. I will try to convey some geometrical intuition for uncertainty, and the relationship between the intrinsic uncertainty of linear inverse problems and the uncertainty of some common techniques applied to them.

References & Acknowledgements

Donoho, D.L., 1994. Statistical Estimation and Optimal Recovery, *Ann. Stat.*, *22*, 238-270.

Evans, S.N. and Stark, P.B., 2002. Inverse Problems as Statistics, *Inverse Problems*, *18*, R1-R43 (in press).

Stark, P.B., 1992. Inference in infinite-dimensional inverse problems: Discretization and duality, *J. Geophys. Res.*, *97*, 14,055-14,082.

Created using TexPoint by G. Necula, http://raw.cs.berkeley.edu/texpoint

Outline

- Inverse Problems as Statistics
  - Ingredients; Models
  - Forward and Inverse Problems—applied perspective
  - Statistical point of view
  - Some connections
- Notation; linear problems; illustration
- Identifiability and uniqueness
  - Sketch of identifiablity and extremal modeling
  - Backus-Gilbert theory
- Decision Theory
  - Decision rules and estimators
  - Comparing decision rules: Loss and Risk
  - Strategies; Bayes/Minimax duality
  - Mean distance error and bias
  - Illustration: Regularization
  - Illustration: Minimax estimation of linear functionals
- Distinguishing Models: metrics and consistency

Inverse Problems as Statistics

- Measurable space X of possible *data*.

- Set Θ of possible descriptions of the world—*models*.
- Family $\mathsf{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$ of probability distributions on X, indexed by models θ.

- *Forward operator* $\theta \mapsto \mathsf{P}_\theta$ maps model θ into a probability measure on X.

Data X are a sample from P_θ.

P_θ is whole story: stochastic variability in the "truth," contamination by measurement error, systematic error, censoring, *etc*.

Models

- Set Θ usually has special structure.
- Θ could be a convex subset of a separable Banach space T. (geomag, seismo, grav, MT, …)
- Physical significance of θ generally gives $\theta \mapsto \mathsf{P}_\theta$ reasonable analytic properties, *e.g.*, continuity.

Forward Problems in Geophysics

Composition of steps:

- transform idealized description of Earth into perfect, noise-free, infinite-dimensional data ("approximate physics")
- censor perfect data to retain only a finite list of numbers, because can only measure, record, and compute with such lists
- possibly corrupt the list with measurement error.

Equivalent to single-step procedure with corruption on par with physics, and mapping incorporating the censoring.

Inverse Problems

Observe data X drawn from distribution $P_?$ for some unknown $\theta \in \Theta$. (Assume Θ contains at least two points; otherwise, data superfluous.)

Use data X and the knowledge that $\theta \in \Theta$ to learn about θ; for example, to estimate a parameter $g(\theta)$ (the value $g(?)$ at ? of a continuous G-valued function g defined on Θ).

Geophysical Inverse Problems

- Inverse problems in geophysics often “solved” using applied math methods for Ill-posed problems (*e.g.*, Tichonov regularization, analytic inversions)
- Those methods are designed to answer different questions; can behave poorly with data (*e.g.*, bad bias & variance)
- Inference ≠ construction: statistical viewpoint more appropriate for interpreting geophysical data.

Elements of the Statistical View

Distinguish between characteristics of the problem, and characteristics of methods used to draw inferences.

One fundamental property of a parameter:

g is *identifiable* if for all $?, \zeta \in T$,

$$\{g(?) \neq g(\zeta)\} \Rightarrow \{P_{\eta} \neq P_{\zeta}\}.$$

In most inverse problems, $g(?) = ?$ not identifiable, and few linear functionals of ? are identifiable.

Deterministic and Statistical Perspectives: Connections

Identifiability—distinct parameter values yield distinct probability distributions for the observables—similar to *uniqueness*—forward operator maps at most one model into the observed data.

Consistency—parameter can be estimated with arbitrary accuracy as the number of data grows—related to *stability* of a recovery algorithm—small changes in the data produce small changes in the recovered model.

$\exists$ quantitative connections too.

More Notation

Let T be a separable Banach space, T^* its normed dual.

Write the pairing between T and T^*

$$\langle \bullet, \bullet \rangle : T^* \times T \to \mathbb{R}.$$

Linear Forward Problems

A forward problem is *linear* if

- T is a subset of a separable Banach space T
- $X = R^n$
- For some fixed sequence $(?_j)_{j=1}^n$ of elements of T^*,

$$X = (X_j)_{j=1}^n, \text{ where}$$

$$X_j = \langle \kappa_j, \theta \rangle + \epsilon_j, \ \theta \in \Theta, \text{ and}$$

$$\epsilon = (\epsilon_j)_{j=1}^n$$

is a vector of stochastic errors whose distribution does not depend on ?.

Linear Forward Problems, contd.

- Linear functionals $\{?_j\}$ are the "representers"
- Distribution $\mathsf{P}_?$ is the probability distribution of

X. Typically, dim(T) = ∞; at least, n < dim(T), so estimating ? is an underdetermined problem.

Define

$$K : T \rightarrow \mathsf{R}^n$$

$$T \mapsto (<?_j, ?>)_{j=1}^{n} .$$

Abbreviate forward problem by X = K? + e, ? ∈ T.

Linear Inverse Problems

Use $X = K? + e$, and the knowledge $? \in T$ to estimate or draw inferences about $g(?)$.

Probability distribution of X depends on ? only through K?, so if there are two points

$$?_1, ?_2 \in T \text{ such that } K?_1 = K?_2 \text{ but}$$

$$g(?_1) \neq g(?_2),$$

then $g(?)$ is not identifiable.

Ex: Sampling w/ systematic and random error

Observe

$$X_j = f(t_j) + \rho_j + \varepsilon_j, \quad j = 1, 2, \ldots, n,$$

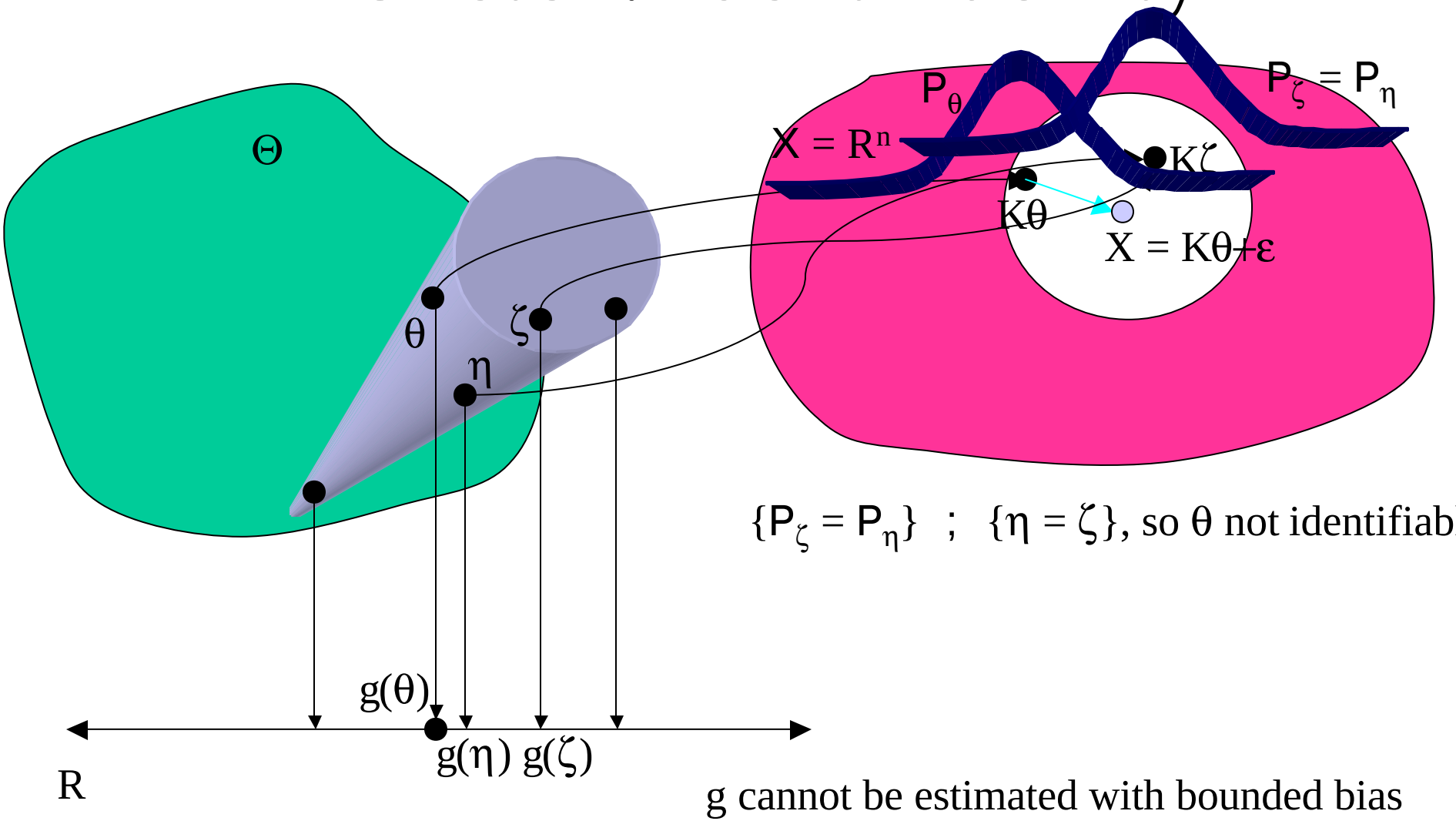
- $f \in C$, a set of smooth of functions on [0, 1]
- $t_j \in [0, 1]$
- $|\rho_j| \leq 1$, $j=1, 2, \ldots, n$
- ε_j *iid* $N(0, 1)$

Take $\Theta = C \times [-1, 1]^n$, $X = R^n$, and $\theta = (f, \rho_1, \ldots, \rho_n)$.

Then P_θ has density

$$(2\pi)^{-n/2} \exp\{-\sum_{j=1}^{n} (x_j - f(t_j)-\rho_j)^2\}.$$

Sketch: Identifiability



$\{P_\zeta = P_\eta\}$; $\{\eta = \zeta\}$, so θ not identifiable

g cannot be estimated with bounded bias

$\{P_\zeta = P_\eta\}$; $\{g(\eta) = g(\zeta)\}$, so g not identifiable

Backus-Gilbert Theory

Let $\Theta = T$ be a Hilbert space.

Let $g \in T = T^*$ be a linear parameter.

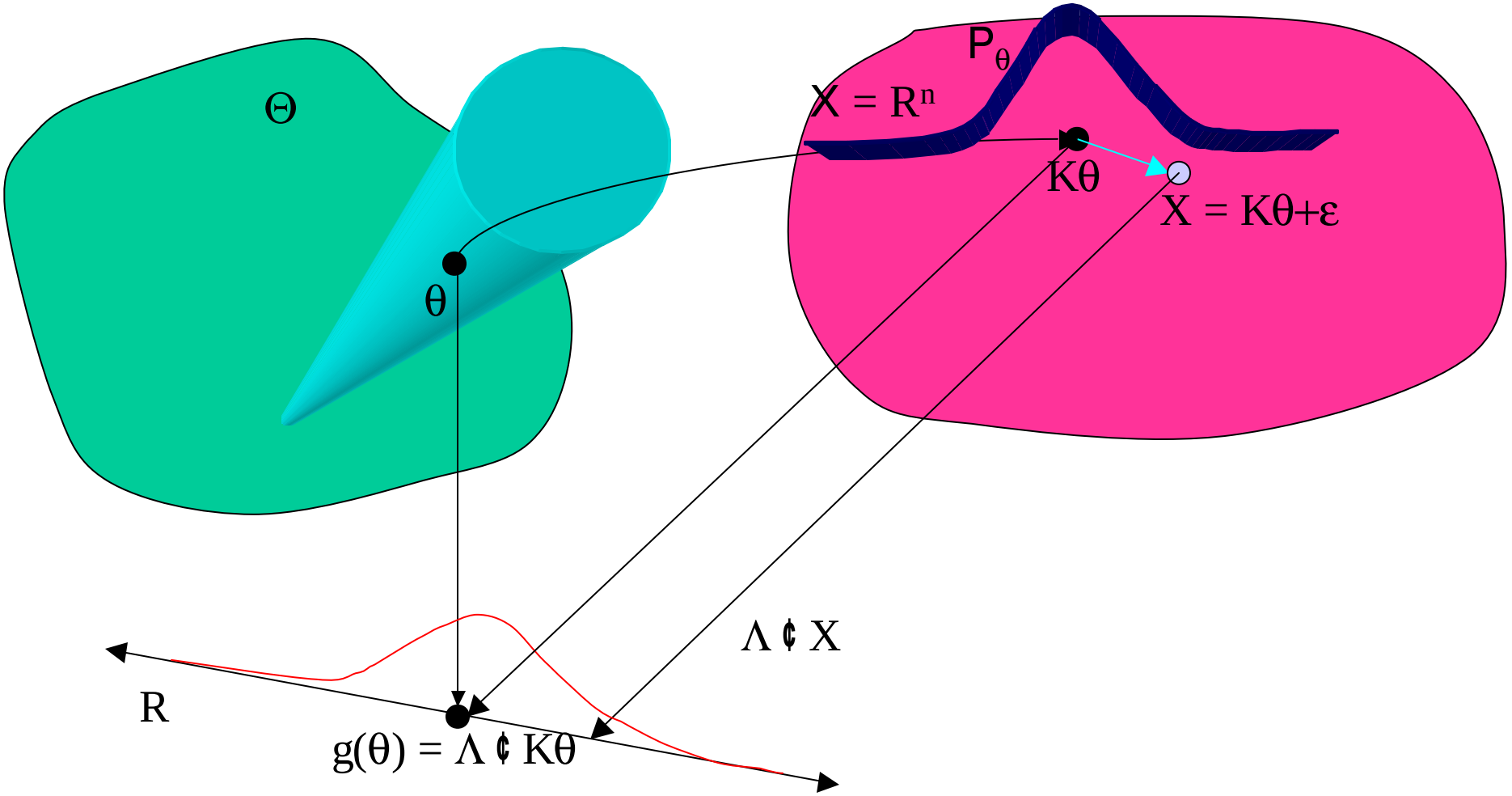
Let $\{\kappa_j\}_{j=1}^{n} \subset T^*$. Then:

$g(\theta)$ is identifiable iff $g = \Lambda \cdot K$ for some $1 \times n$ matrix Λ.

If also $E[\varepsilon] = 0$, then $\Lambda \cdot X$ is unbiased for g.

If also ε has covariance matrix $\Sigma = E[\varepsilon\varepsilon^T]$, then the MSE of $\Lambda \cdot X$ is $\Lambda \cdot \Sigma \cdot \Lambda^T$.

Sketch: Backus-Gilbert



Backus-Gilbert^{++}: Necessary conditions

Let g be an identifiable real-valued parameter. Suppose $\exists\ ?_0 \in T$, a symmetric convex set $T \subseteq \mathcal{T}$, $c \in \mathbb{R}$, and $g: T \to \mathbb{R}$ such that:

1. $?_0 + T \subseteq T$
2. For $t \in T$, $g(?_0 + t) = c + g(t)$, and $g(-t) = -g(t)$
3. $g(a_1t_1 + a_2t_2) = a_1g(t_1) + a_2g(t_2)$, $t_1, t_2 \in T$, $a_1, a_2 \geq 0$, $a_1+a_2 = 1$, and
4. $\sup_{t \in T} |g(t)| < \infty$.

Then $\exists$ 1×n matrix ? s.t. the restriction of g to T is the restriction of ? ·K to T.

Backus-Gilbert^{++}: Sufficient Conditions

Suppose $g = (g_i)_{i=1}^{m}$ is an $\mathbb{R}^m$-valued parameter that can be written as the restriction to T of $? \cdot K$ for some $m \times n$ matrix $?$.

Then

1. g is identifiable.
2. If $E[e] = 0$, $? \cdot X$ is an unbiased estimator of g.
3. If, in addition, e has covariance matrix $S = E[ee^T]$, the covariance matrix of $? \cdot X$ is $? \cdot S \cdot ?^T$ whatever be $\mathbb{P}_?$.

Decision Rules

A (*randomized*) *decision rule*

$$\mathrm{d}: X \rightarrow M_1(A)$$

$$\mathrm{x} \mapsto \mathrm{d}_{\mathrm{x}}(.),$$

is a measurable mapping from the space X of possible data to the collection $M_1(A)$ of probability distributions on a separable metric space A of *actions*.

A *non-randomized decision rule* is a randomized decision rule that, to each $\mathrm{x} \in X$, assigns a unit point mass at some value

$$\mathrm{a} = \mathrm{a}(\mathrm{x}) \in A.$$

Estimators

An *estimator* of a parameter g(?) is a *decision rule* for which the space *A* of possible actions is the space *G* of possible parameter values.

g=g(X) is common notation for an estimator of g(?).

Usually write non-randomized estimator as a *G*-valued function of x instead of a $M_1(G)$-valued function.

Comparing Decision Rules

∃Infinitely many decision rules and estimators.

Which one to use?

The best one!

But what does *best* mean?

Loss and Risk

- 2-player game: Nature v. Statistician.
- Nature picks ? from T.
  ? is secret, but statistician knows T.
- Statistician picks d from a set *D* of rules.
  d is secret.
- Generate data X from $\mathsf{P}_{?}$, apply d.
- Statistician pays *loss* $l(?, d(X))$. l should be dictated by scientific context, but…
- *Risk* is expected loss: $r(?, d) = \mathsf{E}_{\theta} l(?, d(X))$
- Good rule δ has small risk, but what does *small* mean?

Strategy

Rare that one δ has smallest risk $\forall \theta \in \Theta$.

- δ is *admissible* if not dominated.
- *Minimax decision* minimizes $\sup_{\theta \in \Theta} r(?, d)$ over $\delta \in D$
- *Bayes decision* minimizes over $\delta \in D$ for a given *prior probability distribution* π on Θ.

$$\int_{\Theta} r(?,d)p(d?)$$

Minimax is Bayes for *least favorable prior*

Pretty generally for convex Θ, D, concave-convexlike r,

$$\inf_{\delta \in \mathcal{D}} \sup_{\theta \in \Theta} r(\theta, \delta) = \sup_{\pi \in \Pi} \inf_{\delta \in \mathcal{D}} \int_{\Theta} r(\theta, \delta) d\pi(\theta)$$

If minimax risk >> Bayes risk, prior p controls the apparent uncertainty of the Bayes estimate.

Common Risk: Mean Distance Error (MDE)

Let d_G denote the metric on G.

MDE at ? of estimator g *of* g is

$$MDE_{?}(g, g) = \mathsf{E}_{\theta}\,[d(g, g(?))].$$

When metric derives from norm, MDE is called *mean norm error* (MNE).

When the norm is Hilbertian, $(MNE)^2$ is called *mean squared error* (MSE).

Bias

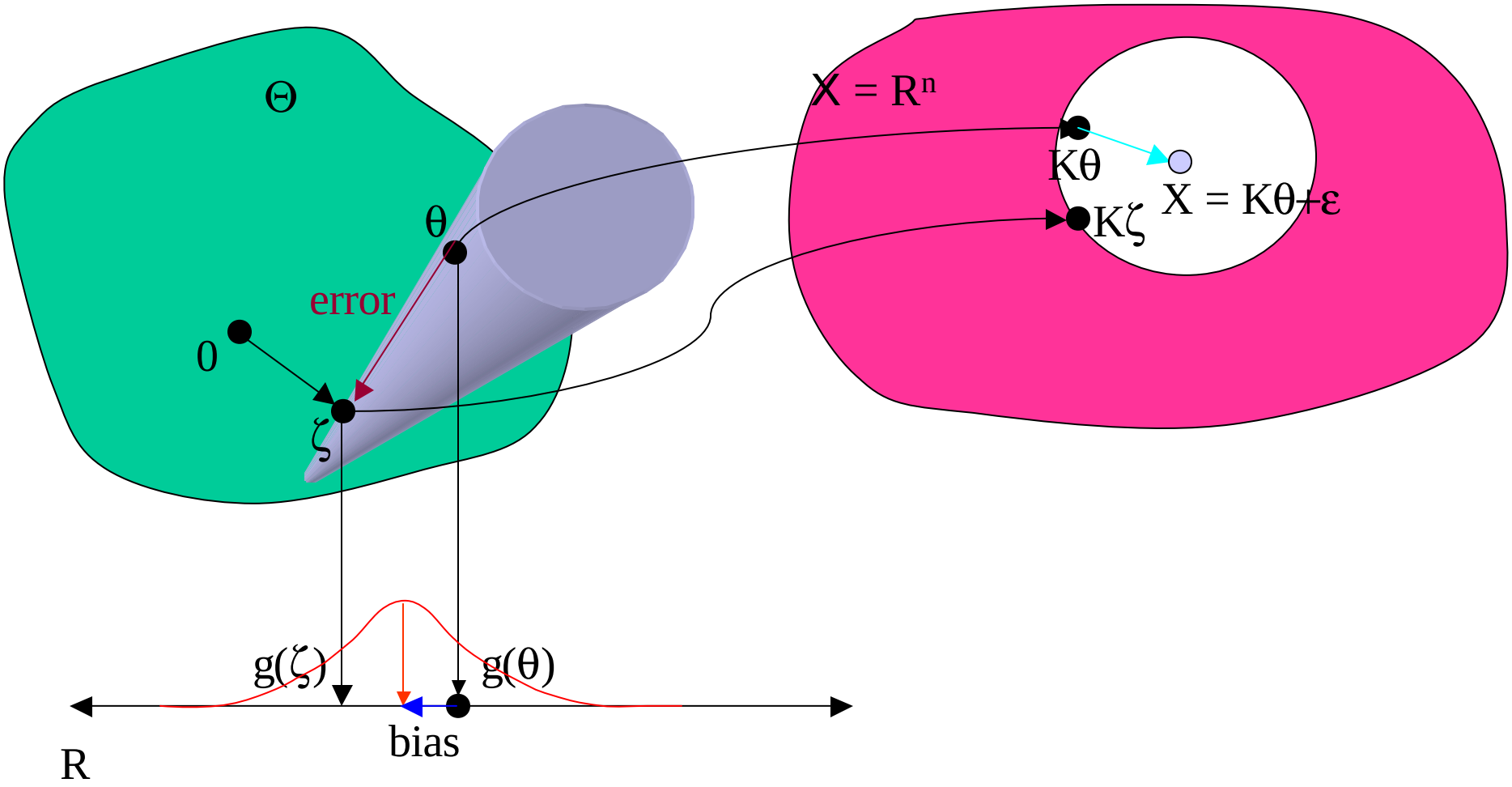
When G is a Banach space, can define *bias at* ? *of* g:

$$bias_?(g, g) = \mathsf{E}_\theta [g - g(?)]$$

(when the expectation is well-defined).

- If $bias_?(g, g) = 0$, say g *is unbiased at* ? (*for* g).
- If g is unbiased at ? for g for every $? \in \Theta$, say g is *unbiased for* g. If such g exists, g is *unbiasedly estimable.*
- If g is *unbiasedly estimable* then g is identifiable.

Sketch: Regularization



$$\zeta = \arg \min_{\{\eta \in \Theta : \|K\eta - X\| \le \chi\}} \|\eta\|$$

Minimax Estimation of Linear parameters: Hilbert Space, Gaussian error

•Observe $X = K\theta + \varepsilon \in \mathbb{R}^n$, with
$\theta \in \Theta \subset T$, T a separable Hilbert space

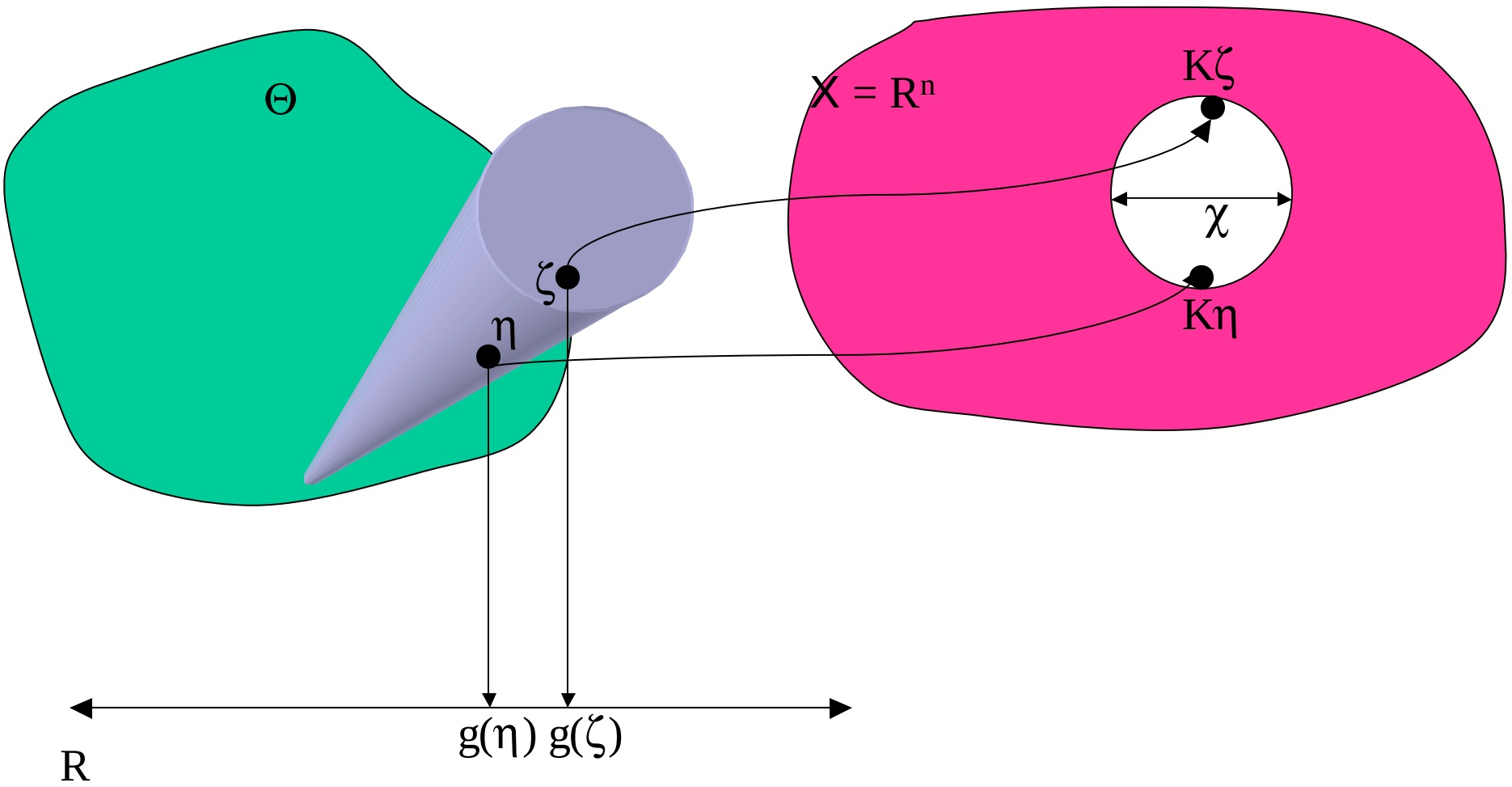
Θ convex
$\{\varepsilon_i\}_{i=1}^n$ *iid* $N(0,\sigma^2)$.

•Seek to learn about $g(\theta)$: $\Theta \to \mathbb{R}$, linear, bounded on Θ

For variety of risks (MSE, MAD, length of fixed-length confidence interval), minimax risk is controlled by *modulus of continuity of g*, calibrated to the noise level.

(Donoho, 1994.)

Modulus of Continuity



$$\rho(g, K, \Theta, \chi) = \sup_{\{\eta,\zeta\in\Theta:\|K\eta-K\zeta\|_2\leq\chi\}} |g(\eta) - g(\zeta)|$$

Distinguishing two models

Data tell the difference between two models ζ and η if the L_1 distance between P_ζ and P_η is large:

$$||\mathbb{P}_\eta - \mathbb{P}_\zeta||_1 = \sup_{|f|\leq 1} \left| \int f d\mathbb{P}_\eta - \int f d\mathbb{P}_\zeta \right|$$

L_1 and Hellinger distances

$$H^2(\mathbb{P},\mathbb{Q}) = \frac{1}{2}\int(\sqrt{d\mathbb{P}} - \sqrt{d\mathbb{Q}})^2.$$

$$\begin{aligned} H^2(\mathbb{P},\mathbb{Q}) &\leq \frac{1}{2}||\mathbb{P}-\mathbb{Q}||_1 \\ &\leq H(\mathbb{P},\mathbb{Q})[||\mathbb{P}+\mathbb{Q}||_1 H^2(\mathbb{P},\mathbb{Q})]^{1/2} \end{aligned}$$

Consistency in Linear Inverse Problems

- $X_i = \kappa_i\theta + \varepsilon_i$, i=1, 2, 3, …
  $\theta \in \Theta$ subset of separable Banach
  $\{\kappa_i\} \subset \Theta^*$ linear, bounded on Θ
  $\{\varepsilon_i\}$ iid μ
- θ *consistently estimable* w.r.t. weak topology iff $\exists\{T_k\}$, T_k Borel function of $X_1, \ldots, X_k$, s.t. $\forall\theta \in \Theta$, $\forall\eta>0$, $\forall\kappa \in \Theta^*$,

$$\lim_k P_\theta\{|\kappa T_k - \kappa\theta| > \eta\} = 0$$

Importance of the Error Distribution

- μ a prob. measure on $\Re$;
  $\mu_a(B) = \mu(B-a)$, $a \in \Re$
- Pseudo-metric on Θ^{**}:

$$D_k(t_1, t_2) = \left\{ \frac{1}{k} \sum_{i=1}^{k} H^2(t_1?_i, t_2?_i) \right\}^{1/2}$$

- If restriction to Θ converges to metric compatible with weak topology, can estimate θ consistently in weak topology.
- For given sequence of functionals $\{\kappa_i\}$, μ rougher $\rightarrow$ consistent estimation easier.

Summary

- Statistical viewpoint is useful abstraction.
  Physics in mapping $\theta \mapsto P_\theta$
  Prior information in constraint $\theta \in \Theta$.
- Separating "model" from parameters of interest is useful: Sabatier's "well posed questions."
- "Solving" inverse problem means different things to different audiences. Thinking about measures of performance is useful.
- Difficulty of problem ≠ performance of specific method.