Statistical inverse problems

Felipe Uribe

Computational Engineering
School of Engineering Sciences
Lappeenranta-Lahti University of Technology (LUT)

Special Course on Inverse Problems
Lappeenranta, FI — January-February, 2024

Recap

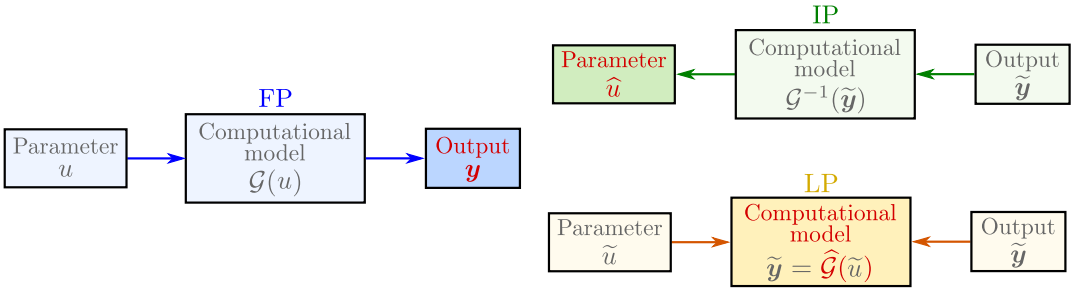


Figure: The first part of the course focused on the blue block. Last part of our course will focus on the Green block.

Why inverse problems?

- The forward problem (FP) is to compute the output, given a system and the input to this system.
- The inverse problem (IP) is to compute the input given the other two quantities (system and output). In most situations, we have noisy measurements of the output.
- Inverse problems are some of the most important mathematical tasks in science and mathematics because they tell us about **parameters that we cannot directly observe**.

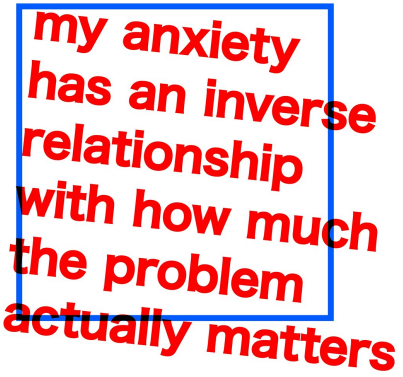


Figure: ;)

PART I: inverse problems

Inverse problems: definition

- Let $y^\dagger \in \mathcal{Y}$ be observational data in some separable Banach space – the data space $\mathcal{Y}$.
- The data will be used to train a mathematical model, that is identify a (true) model parameter $x^\dagger \in \mathcal{X}$. The parameter space $\mathcal{X}$ can also be separable Banach space.
- Let $\mathcal{G} : \mathcal{X} \to \mathcal{Y}$ be a measurable function called the **forward response operator**. It represents the connection between parameter and data in the mathematical model.

Inverse problems: definition

- Let $y^\dagger \in \mathcal{Y}$ be observational data in some separable Banach space – the data space $\mathcal{Y}$.
- The data will be used to train a mathematical model, that is identify a (true) model parameter $x^\dagger \in \mathcal{X}$. The parameter space $\mathcal{X}$ can also be separable Banach space.
- Let $\mathcal{G} : \mathcal{X} \to \mathcal{Y}$ be a measurable function called the forward response operator. It represents the connection between parameter and data in the mathematical model.
- Assuming an *additive observation error*, we define the inverse problem by (see, e.g., [5])

$$\text{find } x^\dagger \in \mathcal{X}, \text{ such that } \; y^\dagger = \mathcal{G}\big(x^\dagger\big) + e^\dagger, \tag{1}$$

where, $e^\dagger \in \mathcal{Y}$ is observational noise. We consider $e^\dagger$ to be unknown and model it as a realization of a random variable with measure ν_{obs}.

Inverse problems: well-posedness

- Inverse problems belong to the class of *ill-posed* problems[1].
- Hadamard's definition says that an inverse problem in well-posed if it satisfies the following three requirements (see, e.g., [2]):
  - **Existence**: The problem must have a solution.
  - **Uniqueness**: There must be only one solution to the problem.
  - **Stability**: The solution must depend continuously on the data. *This means that arbitrarily small perturbations of the data must not produce arbitrarily large perturbations of the solution.*
- If the problem violates one or more of these requirements, it is said to be ill-posed.
- The *existence* condition is in general trivial, the *uniqueness* condition can often be fixed by reformulation of the problem, the *stability* condition is much harder to deal with.

[1] The term was coined in the early 20th century by Jacques S. Hadamard.

Inverse problems: notes

- Note that if the noise takes any value in $\mathcal{Y}$, the inverse problem is ill-posed [5].
- In practice, the inverse problem estimates an x that approximates the ground truth $x^{\dagger}$.
- The noise is oftentimes assumed to be Gaussian distributed with mean zero and non-singular covariance matrix. Other noise assumptions exist in the literature, such as Laplace and Poisson noises.
- The forward response operator can oftentimes be written as $\mathcal{G} = \mathcal{O} \circ G$, defined as the composition of the solution operator G with an observation operator $\mathcal{O}$ that maps the forward solution to the data space.

Discrete inverse problems: classical methods (I)

- In practice, the unknown parameter functions have to be discretized. Hence, the discrete inverse problem will estimate an unknown model parameter $\boldsymbol{x}^\dagger \in \mathcal{X} := \mathbb{R}^d$ using noisy observed data $\boldsymbol{y}^\dagger \in \mathcal{Y} := \mathbb{R}^m$.

- If the forward operator is linear $\mathcal{G}(\boldsymbol{x}) = \mathbf{G}\boldsymbol{x}$ and noise is Gaussian (note similarity with regression), the solution of eq. (1) can be estimated using *ordinary least-squares* (OLS) as

$$\boldsymbol{x}^\dagger \approx \boldsymbol{x}_\alpha = \underset{\boldsymbol{x}\in\mathbb{R}^d}{\arg\min} \frac{1}{2} \left\| \mathbf{G}\boldsymbol{x} - \boldsymbol{y}^\dagger \right\|_2^2 = (\mathbf{G}^\mathsf{T}\mathbf{G})^{-1}\mathbf{G}^\mathsf{T}\boldsymbol{y}^\dagger. \tag{2}$$

- Due to the ill-posedness of eq. (1), OLS might not work in practice, we usually employ deterministic regularization methods based on spectral filtering, such as: truncated SVD, Tikhonov, or iterative algorithms (e.g., Landweber, Kaczmarz).

Discrete inverse problems: classical methods (II)

- A more stable approach uses *Tikhonov regularization* (also known as ridge regression). The potential issue of a near-singular matrix $\mathbf{G}^\mathsf{T}\mathbf{G}$ is alleviated by adding positive elements, thereby decreasing its condition number:

$$\boldsymbol{x}^\dagger \approx \boldsymbol{x}_\alpha = \underset{\boldsymbol{x}\in\mathbb{R}^d}{\arg\min} \frac{1}{2}\left\|\mathbf{G}\boldsymbol{x} - \boldsymbol{y}^\dagger\right\|_2^2 + \frac{\alpha}{2}\left\|\bar{\mathbf{L}}\boldsymbol{x}\right\|_2^2 = (\mathbf{G}^\mathsf{T}\mathbf{G} + \alpha\mathbf{L})^{-1}\mathbf{G}^\mathsf{T}\boldsymbol{y}^\dagger, \tag{3}$$

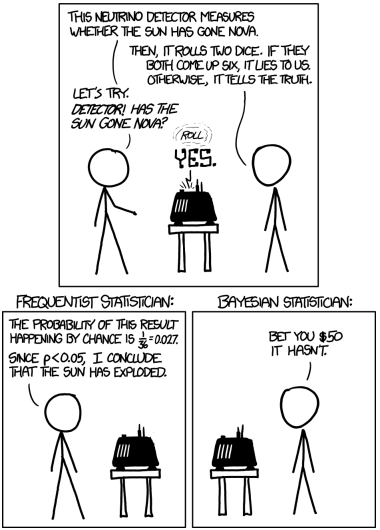
  with $\mathbf{L} = \bar{\mathbf{L}}^\mathsf{T}\bar{\mathbf{L}}$ a regularization matrix, with regularization parameter $\alpha > 0$.

- Regularization can also be achieved using the statistical framework, which also offers a way to model the potential uncertainty about the parameter $\boldsymbol{x}$.

PART II: statistical inverse problems

Statistical inverse problems

- The statistical techniques that we will be most concerned with are based on frequentist and Bayesian methods, and inferences can be drawn from their use [6].
- Did the sun just explode? It is night so we are not sure (from xkcd.com/1132/).
- The Sun gone nova is not repeatable, which makes it highly unsuitable for frequentists which interpret probability as estimate of *how frequent an event is*, given that we can repeat the experiment many times.
- In contrast, Bayesian probability is interpreted as our *degree of belief giving all available prior knowledge*, making it suitable for common sense reasoning about one-time events.



Frequentist inference

- Frequentists do not assign probabilities to the unknown parameters $\boldsymbol{x}$. This means that one can write likelihoods π_{like}, but not priors or posteriors; $\boldsymbol{x}$ is not a random variable.
- The frequentist paradigm considers $\boldsymbol{y}$ resulting from a random and repeatable experiment.
- In the frequentist viewpoint, there is no single preferred methodology for inverting the relationship between parameters and data. Instead, consider various estimators $\widehat{\boldsymbol{x}} \approx \boldsymbol{x}^{\dagger}$.

Frequentist inference

- Frequentists do not assign probabilities to the unknown parameters $\boldsymbol{x}$. This means that one can write likelihoods π_{like}, but not priors or posteriors; $\boldsymbol{x}$ is not a random variable.
- The frequentist paradigm considers $\boldsymbol{y}$ resulting from a random and repeatable experiment.
- In the frequentist viewpoint, there is no single preferred methodology for inverting the relationship between parameters and data. Instead, consider various estimators $\widehat{\boldsymbol{x}} \approx \boldsymbol{x}^{\dagger}$.
- Relies on **hypothesis testing**, bias, mean-square error, confidence intervals to verify the estimator $\widehat{\boldsymbol{x}}$.
- Common frequentist approaches include: (i) Maximum likelihood (ii) BLUE (best linear unbiased estimators), (iii) best asymptotically normal (BAN) estimator, (iv) method of moments estimator (MME), etc.

Frequentist inference: maximum likelihood (I)

- The method of maximum likelihood estimation (MLE) is quite a popular technique for deriving estimators.

- We model a set of observations as a random sample from an unknown joint distribution with density $\pi_{\text{like}}(\boldsymbol{y} \mid \boldsymbol{x})$ which is expressed in terms of a set of parameters $\boldsymbol{x}$. The goal of MLE is to determine the parameters $\boldsymbol{x}$ for which the observed data have the highest joint probability.

Frequentist inference: maximum likelihood (I)

- The method of maximum likelihood estimation (MLE) is quite a popular technique for deriving estimators.
- We model a set of observations as a random sample from an unknown joint distribution with density $\pi_{\text{like}}(\boldsymbol{y} \mid \boldsymbol{x})$ which is expressed in terms of a set of parameters $\boldsymbol{x}$. The goal of MLE is to determine the parameters $\boldsymbol{x}$ for which the observed data have the highest joint probability.
- Evaluating the joint density at the observed data sample $\boldsymbol{y}$ gives a real-valued function

$$\mathcal{L}(\boldsymbol{x}; \boldsymbol{y}) = \pi_{\text{like}}(\boldsymbol{y} \mid \boldsymbol{x}) \propto \exp\left(-\Phi(\boldsymbol{x}; \boldsymbol{y})\right), \tag{4}$$

which is called the likelihood function, and Φ is the negative log-likelihood or potential function.

Frequentist inference: maximum likelihood (II)

- The goal is to find the values of the model parameters that maximize the likelihood function over the parameter space, that is

$$\boldsymbol{x}^\dagger \approx \widehat{\boldsymbol{x}}_{\mathrm{ML}} = \underset{\boldsymbol{x}\in\mathbb{R}^d}{\arg\max}\, \mathcal{L}(\boldsymbol{x};\boldsymbol{y});$$

  intuitively, this selects the parameter value that makes the observed data most probable.

Frequentist inference: maximum likelihood (II)

- The goal is to find the values of the model parameters that maximize the likelihood function over the parameter space, that is

$$\boldsymbol{x}^{\dagger} \approx \widehat{\boldsymbol{x}}_{\mathrm{ML}} = \underset{\boldsymbol{x} \in \mathbb{R}^d}{\arg\max}\, \mathcal{L}(\boldsymbol{x}; \boldsymbol{y});$$

  intuitively, this selects the parameter values that make the observed data most probable.

- In general, no closed-form solution to the maximization problem is available, and an MLE can only be found via numerical optimization. In practice, it is often convenient to work with the potential function.

- As the data size increases to infinity, sequences of MLEs converges in probability to the value being estimated.

- The MLE is identical to solving the inverse problem using OLS.

Frequentist inference: hypothesis testing

- Null-hypothesis significance testing (NHST) is still one of the most dominant approaches to statistical inference, although heavily criticized.

- With NHST we want to test the estimator against the null-hypothesis (i.e., underlying causative relationship does not exist). Moreover, the calculated probability of observing the estimator (or larger), given the null-hypothesis, is called p-value.

- Rather than performing NHST, uncertainty of the estimated parameter can be represented with the confidence interval (CI). **Example**: the 95% CI contains all the hypotheses parameter values that would not be rejected by $p < 0.05$ NHST. This implies that, *in the long-run, 95% CI will capture the true parameter value 95% of the time.*

Bayesian inference: basic timeline

- Thomas Bayes (1763) - problem of inverse probability: *An Essay towards solving a problem in the doctrine of chances*.
- Pierre–Simon Laplace (1774) - *Mémoire sur la probabilité des causes par les événements*.
- Harold Jeffreys (1939) - revival of the "objective" Bayesian view of probability.
- Edwin T. Jaynes (1957) - maximum entropy, Bayesian/information theory.
- Andrew M. Stuart (2010) - foundations in infinite dimensions/inverse problems [7].

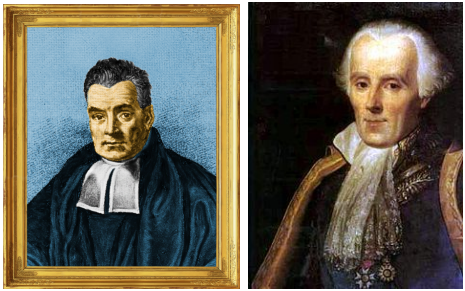


Figure: Bayes (maybe?) and Laplace.

Bayesian inference

- Whereas the difficulties related to MLE methods are mainly *optimization problems*, the Bayesian approach often results in integration problems.

- In the Bayesian paradigm, information brought by the data $\boldsymbol{y}$ (a realization of $\pi_{\text{like}}(\cdot \mid \boldsymbol{x})$), is combined with prior information that is specified in a prior distribution with density $\pi_{\text{pr}}(\boldsymbol{x})$.

- Such information is summarized in a probability density $\pi_{\text{pos}}(\boldsymbol{x} \mid \boldsymbol{y})$, called the posterior. This is derived from the joint density $\pi_{\text{like}}(\boldsymbol{y} \mid \boldsymbol{x})\pi_{\text{pr}}(\boldsymbol{x})$ using Bayes theorem.

Bayesian inverse problem: measure-theoretic (I)

- First, we model the parameter $x \sim \nu_{\mathrm{pr}}$ as a RV. This reflects the uncertainty in the parameter. Moreover, ν_{pr} is the so-called prior measure.

- We assume that x and e are independent RVs defined on an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Therefore, $y := \mathcal{G}(x) + e$ is also a RV, reflecting the distribution of the data, given an uncertain parameter. The conditional measure, given a realized value x', is

$$\nu_{\mathrm{L}} := \mathbb{P}[y \in \cdot \mid x = x'] = \nu_{\mathrm{obs}}(\cdot - \mathcal{G}(x')). \tag{5}$$

- The solution to the Bayesian inverse problem is the posterior measure (given the observed data $y = y^{\dagger}$) [5]

$$\nu_{\mathrm{pos}}^{\dagger} := \mathbb{P}\left[x \in \cdot \;\middle|\; y^{\dagger} = \mathcal{G}(x) + e\right]. \tag{6}$$

Bayesian inverse problem: measure-theoretic (II)

- Bayes' theorem gives a connection of ν_{pr}, ν_{pos} and ν_{L} in terms of their probability densities. The Radon–Nikodym theorem implies that such densities exist:

$$\frac{\mathrm{d}\nu_{\mathrm{L}}}{\mathrm{d}\nu_Y}(y^\dagger) =: \pi_{\mathrm{like}}(y^\dagger \mid x') \qquad \frac{\mathrm{d}\nu_{\mathrm{pr}}}{\mathrm{d}\nu_X}(x) =: \pi_{\mathrm{pr}}(x), \tag{7}$$

  where the dominating measures ν_X, ν_Y are often given by the counting measure, the Lebesgue measure, or a Gaussian measure.

Bayesian inverse problem: measure-theoretic (II)

- Bayes' theorem gives a connection of ν_{pr}, ν_{pos} and ν_{L} in terms of their probability densities. The Radon–Nikodym theorem implies that such densities exist:

$$\frac{\mathrm{d}\nu_{\mathrm{L}}}{\mathrm{d}\nu_Y}(y^\dagger) =: \pi_{\mathrm{like}}(y^\dagger \mid x') \qquad \frac{\mathrm{d}\nu_{\mathrm{pr}}}{\mathrm{d}\nu_X}(x) =: \pi_{\mathrm{pr}}(x), \tag{7}$$

  where the dominating measures ν_X, ν_Y are often given by the counting measure, the Lebesgue measure, or a Gaussian measure.

- The corresponding Radon–Nikodym derivative of the posterior wrt the prior, presents a general version of Bayes' formula [7]

$$\frac{\mathrm{d}\nu^\dagger_{\mathrm{pos}}}{\mathrm{d}\nu_{\mathrm{pr}}}(x) = \frac{1}{Z}\pi_{\mathrm{like}}(y^\dagger \mid x); \tag{8}$$

  for example, if x is infinite-dimensional and ν_{pr} is Gaussian, we set $\nu_X := \nu_{\mathrm{pr}}$ and $\pi_{\mathrm{pr}} \equiv 1$. The posterior measure is then given in terms of a density wrt the Gaussian prior measure.

Bayesian inverse problem: well-posedness

- Stuart [7][2] transferred Hadamard's principle of well-posedness to Bayesian inverse problems: the Bayesian inverse problem is Lipschitz well-posed, if $\nu^{\dagger}_{\mathrm{pos}}$ exists, $\nu^{\dagger}_{\mathrm{pos}}$ is unique, and $y^{\dagger} \to \nu^{\dagger}_{\mathrm{pos}}$ is locally Lipschitz continuous (measured by the Hellinger distance).

- To verify stability, the distance between posteriors is typically measured in the Hellinger distance. However, in practice, it is not possible to show Lipschitz well-posedness for the Bayesian inversion for black-box models. Further, *Hadamard's concept contains only continuity, not Lipschitz continuity.*

- Latz [5][3] extended the notion of well-posedness for a general class of probability metrics, and by considering continuity instead of Lipschitz continuity of the data-to-posterior map.

---

[2] A. M. Stuart. "Inverse problems: a Bayesian perspective". In: *Acta Numerica* 19 (2010), pp. 451–559.

[3] J. Latz. "On the Well-posedness of Bayesian Inverse Problems". In: *SIAM/ASA Journal on Uncertainty Quantification* 8.1 (2020), pp. 451–482.

Bayesian inverse problem: discrete case

- Let us go back to a less abstract setting. Recall that after discretization the unknown parameter is modeled as a random vector $\boldsymbol{X}$ taking values $\boldsymbol{x} \in \mathcal{X} := \mathbb{R}^d$ and the noisy observed data is $\boldsymbol{y}^\dagger \in \mathcal{Y} := \mathbb{R}^m$.

- We define the Bayesian inverse problem (BIP), as the task of characterizing the probability density

$$\pi_{\text{pos}}\left(\boldsymbol{x} \mid \boldsymbol{y}^\dagger\right) = \frac{1}{Z}\pi_{\text{like}}\left(\boldsymbol{y}^\dagger \mid \boldsymbol{x}\right)\pi_{\text{pr}}(\boldsymbol{x}). \tag{9}$$

  - $\pi_{\text{pr}}(\boldsymbol{x})$ is the prior probability density.
  - $\pi_{\text{like}}\left(\boldsymbol{y}^\dagger \mid \boldsymbol{x}\right)$ is the likelihood function.
  - $Z = \int_{\mathbb{R}^d} \pi_{\text{like}}\left(\boldsymbol{y}^\dagger \mid \boldsymbol{x}\right)\pi_{\text{pr}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ is the normalizing constant of the posterior density, called the model evidence[4].

[4] The notation Z follows from the German term *Zustandssumme*.

Bayesian inverse problem: the Gaussian likelihood

- We will assume that the errors are Gaussian with identity correlation matrices, i.e., all the elements of $\boldsymbol{e} \in \mathcal{Y}$ come from the same Gaussian distribution with zero mean and variance σ^2_{obs}. Oftentimes, we use the noise precision $\lambda = 1/\sigma^2_{\text{obs}}$.

- We have seen that the conditional measure of the data given a parameter value follows from the distribution assumed on the noise:

$$\boldsymbol{e} = [\boldsymbol{y} - \mathcal{G}(\boldsymbol{x})] \sim \mathcal{N}(\boldsymbol{e}; \mathbf{0}, \sigma^2_{\text{obs}}\mathbf{I}_m); \tag{10}$$

  due to the additive error assumption, the data misfit $\boldsymbol{y} - \mathcal{G}(\boldsymbol{x})$ follows the noise distribution.

- Then the likelihood function is the conditional probability density of the data given a parameter value, which is just a shifted version of the noise distribution $\mathcal{N}(\boldsymbol{y}; \mathcal{G}(\boldsymbol{x}), \sigma^2_{\text{obs}}\mathbf{I}_m)$:

$$\pi_{\text{like}}\left(\boldsymbol{y}^\dagger \mid \boldsymbol{x}\right) = \mathcal{L}(\boldsymbol{x}; \boldsymbol{y}^\dagger) = \frac{1}{(2\pi)^{m/2}\sigma^m_{\text{obs}}} \exp\left(-\frac{1}{2\sigma^2_{\text{obs}}} \left\|\boldsymbol{y}^\dagger - \mathcal{G}(\boldsymbol{x})\right\|_2^2\right). \tag{11}$$

Point estimators (I)

- The *maximum a posteriori probability* (MAP) estimator (or penalized maximum likelihood, or poor's man Bayesian estimator), which estimates the *mode of the posterior*:

$$\widehat{\boldsymbol{x}}_{\mathrm{MAP}} = \underset{\boldsymbol{x}\in\mathbb{R}^d}{\arg\max} \log(\pi_{\mathrm{pos}}(\boldsymbol{x} \mid \boldsymbol{y})). \tag{12}$$

- The *posterior mean* (PM) (or conditional mean or Bayesian estimator):

$$\widehat{\boldsymbol{x}}_{\mathrm{PM}} = \mathbb{E}_{\pi_{\mathrm{pos}}}[\boldsymbol{x}] = \int_{\mathbb{R}^d} \boldsymbol{x}\, \pi_{\mathrm{pos}}(\boldsymbol{x} \mid \boldsymbol{y})\, \mathrm{d}\boldsymbol{x}. \tag{13}$$

- Posterior credible sets: a set $S_\alpha(\boldsymbol{y}^\dagger) \subset \mathbb{R}^d$, such that $\mathbb{P}\big[\boldsymbol{x} \in S_\alpha(\boldsymbol{y}^\dagger)\big] = 1-\alpha$ is called a posterior $100(1-\alpha)\%$ credible set for $\boldsymbol{x}$.

- **Tip**: highest probability density (HPD) credible sets, median, mode, or MAD, when the underlying posterior is heavy-tailed or multimodal.

Point estimators (II)

- The bias of an estimator $\widehat{\boldsymbol{x}}$ of $\boldsymbol{x}^\dagger$ is defined as

$$\text{Bias}(\widehat{\boldsymbol{x}}) = \mathbb{E}\left[\widehat{\boldsymbol{x}} - \boldsymbol{x}^\dagger\right] = \mathbb{E}[\widehat{\boldsymbol{x}}] - \boldsymbol{x}^\dagger; \tag{14}$$

  the norm of the bias tells us how far $\widehat{\boldsymbol{x}}$ is on average from the true $\boldsymbol{x}^\dagger$.

- If the bias and variance of an estimator exist, the mean squared error (MSE) of the estimator is defined as:

$$\text{MSE}(\widehat{\boldsymbol{x}}) = \mathbb{E}\left[\left\|\widehat{\boldsymbol{x}} - \boldsymbol{x}^\dagger\right\|_2^2\right] = \text{Bias}(\widehat{\boldsymbol{x}}) + \mathbb{V}[\widehat{\boldsymbol{x}}]\,; \tag{15}$$

  it measures the performance of an estimator.

- Finally, we say that an estimator is consistent, if it converges in probability to the true value as the sample size goes to infinity.

Point estimators: Bayes risk I

- Suppose the goal is to estimate the parameter vector $\boldsymbol{x}^\dagger$. We choose and estimator $\widehat{\boldsymbol{x}}(\boldsymbol{y}^\dagger) \approx \boldsymbol{x}^\dagger$ and a squared-error *loss function* to compare them:

$$f_2\left(\widehat{\boldsymbol{x}}, \boldsymbol{x}^\dagger\right) = \left\|\widehat{\boldsymbol{x}}(\boldsymbol{y}^\dagger) - \boldsymbol{x}^\dagger\right\|_2^2. \tag{16}$$

- The expected value of the squared-error loss is the MSE of the estimator:

$$\mathsf{MSE}(\widehat{\boldsymbol{x}}) = \mathbb{E}\left[f_2\left(\widehat{\boldsymbol{x}}(\boldsymbol{y}^\dagger), \boldsymbol{x}^\dagger\right)\right]. \tag{17}$$

- Note that the mean of the loss function depends on the unknown value $\boldsymbol{x}^\dagger$. To obtain an overall measure of performance of the estimator, we impose a prior distribution π_{pr} on $\boldsymbol{x}$.

Point estimators: Bayes risk II

- The Bayes risk of $\widehat{\boldsymbol{x}}(\boldsymbol{y}^\dagger)$, for a loss function f and prior distribution π_{pr}, is defined as:

$$R(\widehat{\boldsymbol{x}}) = \mathbb{E}\left[f\left(\widehat{\boldsymbol{x}}(\boldsymbol{y}^\dagger), \boldsymbol{x}\right)\right] \tag{18a}$$

$$= \mathbb{E}_\theta\left[\mathbb{E}_{y|x}\left[f\left(\widehat{\boldsymbol{x}}(\boldsymbol{y}^\dagger), \boldsymbol{x}\right) \mid \boldsymbol{x}\right]\right]. \tag{18b}$$

- In plain terms, the Bayes risk is the **average MSE**. Particularly, it can be seen as the *loss averaged over the parameter and the data*.

- An estimator that minimizes the Bayes risk is called a Bayesian estimator. The posterior mean is the minimizer of the Bayes risk, for any prior and likelihood, with respect to the squared loss (finite variance).

PART III: The linear Gaussian case

Gaussian algebra

- There are many ways to motivate the prevalence of the Gaussian distribution. It is sometimes presented as arising from analytic results like the CLT, ...
- ...or the fact that the Gaussian distribution is the unique probability distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ maximizing the differential entropy functional (next Lecture).
- But the primary practical reason for the ubiquity of Gaussian probability distributions is that they have convenient algebraic properties.
- This is analogous to the popularity of linear approximations in numerical computations: The main reason to construct linear approximations is that linear functions offer a rich analytic theory, and that computers are good at the basic linear operations — addition and multiplication [3].

Gaussian algebra

- In fact, the connection between linear functions and Gaussian distributions runs deeper: Gaussians are a family of probability distributions that are preserved under all linear operations.
- The following properties will be used extensively:
  - If a RV $\boldsymbol{X}$ is Gaussian distributed, then every affine transformation of it also has a Gaussian:

$$\pi_{\boldsymbol{X}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \text{and} \quad \boldsymbol{Y} = \mathbf{G}\boldsymbol{X} + \boldsymbol{b}, \quad \text{then} \quad \pi_{\boldsymbol{Y}}(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}; \mathbf{G}\boldsymbol{\mu} + \boldsymbol{b}, \mathbf{G}\boldsymbol{\Sigma}\mathbf{G}^{\mathsf{T}}).$$

  - The product of two Gaussian density functions is another Gaussian, scaled by a constant:

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}^\star, \boldsymbol{\Sigma}^\star)\mathcal{N}(\boldsymbol{\mu}_1; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2).$$

    where $\boldsymbol{\Sigma}^\star = (\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})^{-1}$ and $\boldsymbol{\mu}^\star = \boldsymbol{\Sigma}^\star \left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right)$.

- These two properties also provide the mechanism for Gaussian inference as we see next.

Linear Gaussian BIPs

Conjugate prior for a Gaussian linear model (with system matrix $\mathbf{G} \in \mathbb{R}^{d \times d}$):

- If the prior density is Gaussian $\pi_{\text{pr}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\text{pr}}, \boldsymbol{\Sigma}_{\text{pr}})$.
- And the likelihood is also Gaussian $\pi_{\text{like}}(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}; \mathbf{G}\boldsymbol{x}, \boldsymbol{\Sigma}_{\text{obs}})$.
- Then the posterior is also Gaussian $\pi_{\text{pos}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\text{pos}}, \boldsymbol{\Sigma}_{\text{pos}})$, with parameters[5] [6]:
  (i) Version 1:

$$\boldsymbol{\Sigma}_{\text{pos}} = \boldsymbol{\Sigma}_{\text{pr}} - \mathbf{C}\mathbf{G}\boldsymbol{\Sigma}_{\text{pr}} \qquad \boldsymbol{\mu}_{\text{pos}}(\boldsymbol{y}) = \boldsymbol{\mu}_{\text{pr}} + \mathbf{C}(\boldsymbol{y} - \mathbf{G}\boldsymbol{\mu}_{\text{pr}}), \tag{19}$$

  where $\mathbf{C} = \boldsymbol{\Sigma}_{\text{pr}}\mathbf{G}^{\mathsf{T}} \left(\mathbf{G}\boldsymbol{\Sigma}_{\text{pr}}\mathbf{G}^{\mathsf{T}} + \boldsymbol{\Sigma}_{\text{obs}}\right)^{-1}$.

  (ii) Version 2:

$$\boldsymbol{\Sigma}_{\text{pos}} = \left(\boldsymbol{\Sigma}_{\text{pr}}^{-1} + \mathbf{G}^{\mathsf{T}}\boldsymbol{\Sigma}_{\text{obs}}^{-1}\mathbf{G}\right)^{-1} \qquad \boldsymbol{\mu}_{\text{pos}}(\boldsymbol{y}) = \boldsymbol{\Sigma}_{\text{pos}} \left(\mathbf{G}^{\mathsf{T}}\boldsymbol{\Sigma}_{\text{obs}}^{-1}\boldsymbol{y} + \boldsymbol{\Sigma}_{\text{pr}}^{-1}\boldsymbol{\mu}_{\text{pr}}\right). \tag{20}$$

[5] The derivation can be consulted in [4, p. 78]

[6] I highly recommend the usage of the Python library `sksparse.cholmod` (Link) for sparse computations when working with high-dimensional Gaussians.

Sampling a Gaussian posterior using optimization (I)

- Find the posterior distribution involves the inversion of some potentially large matrices. These can turn the problem infeasible in practice.
- The most direct sampling algorithm for a Gaussian distribution is based on the Cholesky factorization. In this case, a sample from the Gaussian posterior is obtained as

$$\boldsymbol{x}^{\star} = \boldsymbol{\mu}_{\text{pos}} + \boldsymbol{\Lambda}_{\text{pos}}^{-1/2}\boldsymbol{z},$$

  where $\boldsymbol{\Lambda}_{\text{pos}} = \boldsymbol{\Sigma}_{\text{pos}}^{-1}$ is the precision matrix, $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ is a standard Gaussian random vector, and $\boldsymbol{\Lambda}_{\text{pos}}^{1/2}$ is a lower triangular matrix with real and positive diagonal entries (Cholesky factor).
- We can reformulate the problem starting from the standard Gaussian sampling formula, going to the so-called normal equations and finally writing its least-squares form.

Sampling a Gaussian posterior using optimization (II)

- Replacing eq. (20) into the Gaussian sampling formula using precision matrices instead of covariances, and based on the fact that our noise precision matrix is $\lambda \mathbf{I}_m$, we obtain:

$$\boldsymbol{x}^\star = \boldsymbol{\Lambda}_{\text{pos}}^{-1} \left(\lambda \, \mathbf{G}^\mathsf{T} \boldsymbol{y} + \boldsymbol{\Lambda}_{\text{pr}} \boldsymbol{\mu}_{\text{pr}} \right) + \boldsymbol{\Lambda}_{\text{pos}}^{-1/2} \boldsymbol{z}. \tag{21}$$

- Multiplying both sides by $\boldsymbol{\Lambda}_{\text{pos}}$, we obtain

$$\boldsymbol{\Lambda}_{\text{pos}} \boldsymbol{x}^\star = \left(\lambda \, \mathbf{G}^\mathsf{T} \boldsymbol{y} + \boldsymbol{\Lambda}_{\text{pr}} \boldsymbol{\mu}_{\text{pr}} \right) + \boldsymbol{\Lambda}_{\text{pos}}^{1/2} \boldsymbol{z} \tag{22a}$$

$$\left(\boldsymbol{\Lambda}_{\text{pr}} + \lambda \mathbf{G}^\mathsf{T} \mathbf{G} \right) \boldsymbol{x}^\star = \left(\lambda \, \mathbf{G}^\mathsf{T} \boldsymbol{y} + \boldsymbol{\Lambda}_{\text{pr}} \boldsymbol{\mu}_{\text{pr}} \right) + \left(\boldsymbol{\Lambda}_{\text{pr}} + \lambda \mathbf{G}^\mathsf{T} \mathbf{G} \right)^{1/2} \boldsymbol{z}. \tag{22b}$$

- Working out the expression for the case $\boldsymbol{\mu}_{\text{pr}} = \mathbf{0}$ yields a *perturbed version* of the so-called normal equations which are solved for $\boldsymbol{x}^\star$.

Sampling a Gaussian posterior using optimization (III)

- From the normal equations, the task of sampling a Gaussian random vector can be written as a least-squares problem. We draw a sample $\boldsymbol{x}^\star$ from the posterior by solving (assuming $\boldsymbol{\mu}_{\mathrm{pr}} = \mathbf{0}$):

$$\boldsymbol{x}^\star = \underset{\boldsymbol{x}\in\mathbb{R}^d}{\arg\min} \|\mathbf{M}\boldsymbol{x} - \boldsymbol{z}\|_2^2 \quad \text{with} \quad \mathbf{M} = \begin{bmatrix} \lambda^{1/2}\mathbf{G} \\ \delta^{1/2}\mathbf{L}_{\mathrm{sq}} \end{bmatrix}, \quad \boldsymbol{z} = \begin{bmatrix} \lambda^{1/2}\boldsymbol{y} \\ \mathbf{0}_d \end{bmatrix} + \widetilde{\boldsymbol{z}}, \tag{23}$$

  where (assuming constant prior variance) $\delta = 1/\sigma_{\mathrm{pr}}^2$ is a prior precision parameter, $\mathbf{L}_{\mathrm{sq}}$ is a square-root of the prior structure matrix[7], and $\widetilde{\boldsymbol{z}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{m+d})$. Here, we can use the scipy function `optimize.least_squares(lambda x: M(x)-z, x0)`.

- Nonlinear least-squares can be used when the forward operator is nonlinear. In this case, we can use the Levenberg–Marquardt to solve the least-squares task (see, e.g., [1, p.118]).

[7] We call a *structure matrix* $\mathbf{L}$ to the inverse of the correlation matrix $\mathbf{R}$; which is analogous to the *precision matrix* being the inverse of the covariance matrix. Note that for constant variance, $\boldsymbol{\Sigma}_{\mathrm{pr}} = \sigma_{\mathrm{pr}}^2\mathbf{R}$ and $\boldsymbol{\Lambda}_{\mathrm{pr}} = \delta\mathbf{L}$, with $\mathbf{L} = \mathbf{L}_{\mathrm{sq}}^{\mathsf{T}}\mathbf{L}_{\mathrm{sq}}$.

Final comments (I)

- The posterior distribution can be correlated, even if the prior is uncorrelated.
- Since marginalization (sum rule) and conditioning (product rule) are the two elementary operations of probability theory, "Gaussian distributions map probability theory to linear algebra" — to matrix multiplication and inversion.
- The task of sampling a Gaussian can be posed as a least-squares problem. Then we can use efficient optimization methods to draw samples from high-dimensional Gaussian distributions. For example, the *conjugate gradient* method.

Final comments (II)

- We have seen that statistical inverse problems can be approached from a frequentist (optimization) perspective, or from a Bayesian perspective (integration).
- Oftentimes, computing the posterior distributions is a complicated task. We can rely on approximation methods to approach the problems in a simplified manner.
- **Gaussian densities provide a link between probabilistic inference and linear algebra**. Though of limited expressiveness, they thus form the basis for computationally efficient inference [3].
- The parameters of Gaussian models can be inferred using hierarchical inference (next lecture). In most cases this poses a nonlinear (non-Gaussian) optimization/inference problem. But in the special cases, conjugate priors allow analytic inference.

References

[1] J. M. Bardsley. *Computational Uncertainty Quantification for Inverse Problems*. Society for Industrial and Applied Mathematics (SIAM), 2019.

[2] P. C. Hansen. *Discrete Inverse Problems: Insight and Algorithms*. Society for Industrial and Applied Mathematics (SIAM), 2010.

[3] P. Hennig et al. *Probabilistic Numerics: Computation as Machine Learning*. Cambridge University Press, 2022.

[4] J. Kaipio et al. *Statistical and computational inverse problems*. Springer, 2005.

[5] J. Latz. "On the Well-posedness of Bayesian Inverse Problems". In: *SIAM/ASA Journal on Uncertainty Quantification* 8.1 (2020), pp. 451–482.

[6] C. P. Robert et al. *Monte Carlo statistical methods*. 2nd ed. Springer, 2004.

[7] A. M. Stuart. "Inverse problems: a Bayesian perspective". In: *Acta Numerica* 19 (2010), pp. 451–559.