

INVERSE PROBLEMS AND LEARNING

ÉCOLE CENTRALE CASABLANCA · 2026

Objectives and probabilistic modelling

C1 · Graduate course

Hermann Agossou^{1,2}

PhD candidate, UTER MID@S, ECC / UH2C

¹École Centrale Casablanca · ²Université Hassan II de Casablanca

PART 01

The vocabulary

01

THE VOCABULARY

What does your training loop minimize

Two questions before we start

01

Your last objective

Name the loss you used most recently.
Why that one, and who decided?

02

The counterfactual

Had you chosen differently, what would
have changed about the claim you made?

KEEP YOUR ANSWER TO THE SECOND ONE

Most answers are about convenience or convention. By the end of the session that question has a precise answer, and it is not about convenience.

Three teams, three losses, one dataset

They may be answering different questions

01

Team A: squared error

Targets a conditional mean under squared decision loss. A Gaussian likelihood is one stronger reading.

02

Team B: absolute error

Targets a conditional median under absolute decision loss. Robustness is a separate empirical question.

03

Team C: Poisson NLL

Can encode a count-data likelihood when its support, rate, and dependence assumptions are actually stated.

THREE LEGITIMATE OPTIMIZATION PROBLEMS, THREE DIFFERENT COMMITMENTS

The objective names what is optimized. It does not, by itself, identify a data-generating model or prove that the chosen decision target matches the application.

Probability, likelihood, density

Same expression, different variable held fixed

THE SAME EXPRESSION, READ IN TWO DIRECTIONS

one expression, two readings

Write $p(D \mid \theta)$. Fix θ and vary D : this is a mass or density over possible data. Fix the observed D and vary θ : this is a likelihood. The formula does not change; its domain and normalization obligation do.

$$p(D \mid \theta) \text{ with } \theta \text{ fixed} \quad \longleftrightarrow \quad L(\theta; D) = p(D \mid \theta) \text{ with } D \text{ fixed}$$

The symbolic version. Same right-hand side; only the variable being varied differs.

NOTATION FOR THIS TRACK, STATED ONCE

σ is reserved for the unknown physical parameter in tracks B and D. Here the Gaussian noise scale is σ_n , with $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$; the Poisson rate is r ; the Bernoulli probability is π . The Gaussian mean is written $f_\theta(x)$ because it is the model output.

Density, mass, and what p of x means

The notation everything else is written in

DEFINITION: MASS AND DENSITY

discrete case / continuous case

Discrete: $p(x)$ is the probability of that exact outcome, in $[0,1]$, summing to one. Continuous: the probability of any exact value is zero, so $p(x)$ is a density; probability appears only on integrating. A density is not a probability and may exceed one.

discrete: $\sum_x p(x) = 1$ · continuous: $\int p(x) dx = 1$, while $p(x)$ itself may exceed 1

$$P(a \leq Y \leq b) = \int_a^b p(y) dy \quad \text{while} \quad P(Y = y) = 0 \text{ for every } y$$

Both hold at once for a continuous variable, and neither contradicts the other.

THE COUNTEREXAMPLE WORTH REMEMBERING

A Gaussian with $\sigma_n = 0.01$ has density about 40 at its peak. Nothing is wrong: it is large because it is concentrated on a narrow interval, and its integral is still one.

Conditionally independent is not the same as i.i.d.

The assumption that licenses a product

DEFINITION: CONDITIONAL INDEPENDENCE

conditional, not absolute

Given θ , one observation provides no further information about another. They still share θ ; only the remaining randomness factorizes.

i.i.d. adds one shared law • unequal s_i can be independent without being identical

$$p(y_1, \dots, y_N \mid X, \theta) = \prod_{i=1}^N p(y_i \mid x_i, \theta) \quad \implies \quad \log p = \sum_{i=1}^N \log p(y_i \mid x_i, \theta)$$

The logarithm turns the licensed product into the sum optimized in code.

REMEMBER THAT THIS IS AN ASSUMPTION

Nothing forces this factorization. It is a claim about the data, and part 4 deliberately removes it.

THE VOCABULARY

Casella and Berger: the likelihood function

A data density is not parameter probability

LIKELIHOOD IS NOT A POSTERIOR DENSITY

Statistical Inference, 2nd ed.

The likelihood is the sampling mass or density read as a function of the parameter with data fixed. It has no normalization obligation over θ . A bare integral over θ depends on a chosen measure and parameterization; marginal likelihood instead integrates against a prior.

$L(\theta; D)$ need not integrate to one over θ

See: [1] Casella & Berger (2002)

This is not pedantry. It is the difference between maximum likelihood and Bayesian inference, and it decides what your fitted number is allowed to be called.

PART 02

From model to loss

02

The chain from model to objective

Predict, distribute, multiply, log, negate

WORKFLOW 01

Predict

The model outputs $f_{\theta}(x)$.

WORKFLOW 04

Log

Turn licensed products into sums.

WORKFLOW 02

Distribute

Say which distribution $f_{\theta}(x)$ parameterises.

WORKFLOW 03

Factor

Build the joint likelihood under a stated dependence model.

WORKFLOW 05

Negate

Minimize the negative log-likelihood.

$$\mathcal{L}_{\text{NLL}}(\theta) = - \sum_i \log p(y_i \mid x_i, \theta)$$

Maximising the likelihood and minimising this are the same problem: the logarithm is increasing, and the minus sign flips the direction.

FROM MODEL TO LOSS

Unequal scales produce weighted square loss

Fully derived, then checked numerically

$$\begin{aligned} Y_i \mid x_i, \theta &\sim \mathcal{N}(f_\theta(x_i), s_i^2), \quad s_i > 0 \text{ known}, \\ \ell_i(\theta) &= \frac{1}{2} \left(\frac{y_i - f_\theta(x_i)}{s_i} \right)^2 + \log s_i + \frac{1}{2} \log(2\pi), \\ \mathcal{L}_{\text{NLL}}(\theta) &= \ell_1(\theta) + \cdots + \ell_N(\theta) \end{aligned}$$

VERIFICATION

Sum scalar negative log-densities and compare with the vectorized NLL. A mismatch exposes a lost scale or normalization term.

FROM MODEL TO LOSS

Finish the Bernoulli derivation

The model is given, the algebra is yours

$$y \mid x, \theta \sim \text{Bernoulli}(\pi_\theta(x)) : \quad p(y \mid x, \theta) = \pi_\theta(x)^y (1 - \pi_\theta(x))^{1-y}, \quad y \in \{0, 1\}$$

Given. The exponents simply select π when $y = 1$ and $1 - \pi$ when $y = 0$.

TO COMPLETE

5 minutes. Artifact: one derivation per pair from the Bernoulli mass to binary cross-entropy, with the conditional-independence step boxed. Check it against a second pair before the reveal.

FROM MODEL TO LOSS

Categorical gives you cross-entropy

The multiclass case, derived in full

$$\ell_i(\theta) = -y_i^\top \log p_\theta(x_i), \quad \mathbf{1}^\top y_i = \mathbf{1}^\top p_\theta(x_i) = 1, \quad \mathcal{L}(\theta) = \ell_1(\theta) + \cdots + \ell_N(\theta)$$

VERIFICATION

The one-hot product selects the true class. Compare the final sum with library cross-entropy on the same logits and targets.

Poisson gives you the Poisson NLL

Counts, and the term you are allowed to drop

$$\begin{aligned}p(y \mid x, \theta) &= r_{\theta}(x)^y e^{-r_{\theta}(x)} / y!, \quad y \in \{0, 1, \dots\}, \quad r_{\theta}(x) > 0, \\ \ell_i(\theta) &= r_{\theta}(x_i) - y_i \log r_{\theta}(x_i) + \log(y_i!), \\ \mathcal{L}_{\text{NLL}}(\theta) &= \ell_1(\theta) + \dots + \ell_N(\theta)\end{aligned}$$

VERIFICATION

The rate must stay positive. Compute the objective with and without $\log y!$; the minimizer must be unchanged because that term is data-only.

FROM MODEL TO LOSS

The assumption ledger for the assessed Gaussian case

Six entries, none optional

Layer	Declared assumption	If changed
Support	$Y_i \in \mathbb{R}$	Counts or classes require a different law
Measure	Lebesgue density	A discrete outcome uses probability mass
Conditional law	Gaussian around $f_\theta(x_i)$	Tail and support commitments change
Parameterization	$f_\theta(x_i)$ is the conditional mean	A different target changes the interpretation
Dependence	Independent conditional on X, θ	Use a joint covariance, not a product
Scale	Known $s_i > 0$, allowed to differ	Unknown or predicted scales add parameter terms

FROM MODEL TO LOSS

Bishop: the standard losses and what they assume

Including where weight decay comes from

THREE RESULTS, ONE SOURCE

PRML, sections 1.2.5, 3.1, and 4.3

- ▶ Sum-of-squares error arises from maximum likelihood under a Gaussian noise model of constant variance.
- ▶ Cross-entropy errors arise the same way from Bernoulli and multinomial models.
- ▶ Under MAP estimation, a zero-mean isotropic Gaussian prior contributes a quadratic penalty with scale fixed by its variance.

See: [2] Bishop (2006)

PART 03

Reading it backwards

03

What MSE does and does not identify

One objective, several compatible readings

01

Decision-theoretic

When the conditional second moment exists, expected squared loss is minimized by the conditional mean. No Gaussian assumption is needed.

02

Likelihood

Gaussian, conditionally independent, constant-variance errors make empirical squared error an NLL up to scale and constants.

03

Surrogate

An optimizer may use squared loss for tractability without claiming that it is a log-likelihood or the final decision cost.

The objective alone does not choose among these readings. Additional assumptions and the decision target do.

Compare: [3] Miller et al. (1993); [2] Bishop (2006)

Audit the same squared objective in reverse

One expression, at least two licensed readings

$$J(\theta) = \sum_i (y_i - f_\theta(x_i))^2, \quad \mathcal{L}_{\text{NLL}}(\theta) = \frac{J(\theta)}{2s^2} + N \log(\sqrt{2\pi} s)$$

Likelihood reading: conditionally independent Gaussian errors with one fixed scale make the NLL an affine transform of J.

TWO LICENSED READINGS

- ▶ Likelihood: real response, Gaussian law, conditional independence, fixed common scale
- ▶ Decision: finite conditional second moment; target is the conditional mean

VERIFICATION: A NON-GAUSSIAN COUNTEREXAMPLE

Masses 0.8, 0.1, 0.1 on values 0, 1, 9 give mean 1. Squared risk is 7.2 at decision 1 versus 8.2 at decision 0, with no Gaussian law.

READING IT BACKWARDS

Your turn: audit an objective you use

Multiple readings, defended aloud

TASK

12 minutes. Artifact: a three-column audit table with at least two compatible interpretations, each interpretation's added assumption, and the evidence that could support or refute it.

- ▶ Separate likelihood, decision-theoretic, and surrogate readings
- ▶ Name the added assumption before checking it against the data
- ▶ Check: a randomly chosen peer must challenge one unsupported inference
- ▶ Do not turn algebraic compatibility into a model-adequacy claim

READING IT BACKWARDS

Five-minute break

Resume with one question

WHEN YOU RETURN

What did your objective fail to identify without extra assumptions?

Likelihood, posterior, and where priors enter

MLE, MAP, and the regularizer that was a prior

DEFINITION: MAXIMUM A POSTERIORI

MLE plus a prior

MLE asks which θ makes the observed data most plausible. MAP asks the same question after multiplying in a prior belief about θ itself. Taking logs turns that product into a sum: your data term, plus a penalty.

τ = prior standard deviation on each parameter • small τ = a strong belief that θ is near zero

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} [\log p(D \mid \theta) + \log p(\theta)], \quad p(\theta) = \mathcal{N}(0, \tau^2 I) \implies +\lambda \|\theta\|^2, \quad \lambda = \frac{1}{2\tau^2}$$

Under this chosen isotropic Gaussian prior, the negative log-prior is quadratic and λ records its strength.

PART 04

When the simple picture breaks

Latent variables and the complete data

What EM is actually completing

DEFINITION: LATENT VARIABLE AND COMPLETE DATA

what is unobserved

Sometimes each observation carries a hidden label you never see: which mixture component produced this point. That label is the latent variable. If someone handed you all the labels, fitting would be easy. That imagined dataset is the complete data. EM works by averaging over the labels you do not have.

C_i = hidden component label of observation i · (y_i, C_i) = complete data · γ_{ij} = responsibility

$$p(y \mid \theta) = \sum_C p(y, C \mid \theta), \quad \gamma_{ij} = P(C_i = j \mid y_i, \theta)$$

The observed likelihood is the complete one summed over labels. γ is the responsibility, and is what the E step computes.

WHEN THE SIMPLE PICTURE BREAKS

Dempster, Laird and Rubin: the EM algorithm

What the 1977 paper actually establishes

WORKFLOW 01

E step

Compute the responsibilities γ under the current θ , then form the expected complete-data log-likelihood.

WORKFLOW 02

M step

Exact EM maximises that expectation over θ ; generalized EM need only increase it.

WORKFLOW 03

Repeat

Each iteration does not decrease the observed-data likelihood.

THEOREM: EM LIKELIHOOD MONOTONICITY

JRSS-B, 1977

An exact EM step, or a generalized EM step that increases Q , produces an observed-data likelihood at least as large as before. Parameter convergence and optimality require separate conditions.

$\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)})$ for every t : an inequality, not a limit

See: [4] Dempster et al. (1977)

What E and M actually guarantee

Monotone is not the same as convergent

WHAT IS GUARANTEED

By the 1977 paper.

- ▶ Exact EM maximises Q; generalized EM increases it
- ▶ Observed-data likelihood does not decrease

WHAT IS NOT

And what students assert anyway.

- ▶ Convergence to the maximum likelihood estimate
- ▶ Convergence to a global optimum
- ▶ Independence from the initialisation

A BOUNDED INCREASING VALUE CONVERGES, BUT TO WHAT?

Convergence of likelihood values does not imply parameter convergence. Wu obtains stationary limit points under added compactness and continuity, and the unique MLE only under stronger conditions.

WHEN THE SIMPLE PICTURE BREAKS

EM on one declared mixture: three monotone destinations

The trace is evidence, not a global certificate

initial means	initial log L	final means	final log L	all increments
(-4, 0)	-50.188	(-4.000, 2.000)	-30.928	non-negative
(0, 4)	-50.188	(-2.000, 4.000)	-30.928	non-negative
(-4, 4)	-46.837	(-2.697, 2.697)	-34.976	non-negative

VERIFICATION BEFORE INTERPRETATION

Recompute each initial likelihood, then check every `np.diff(trace)`. A decrease falsifies the implementation contract; three increasing traces do not certify parameter convergence or a global optimum.

WHEN THE SIMPLE PICTURE BREAKS

The covariance matrix of a Gaussian process

Where the kernel enters, and what it encodes

DEFINITION: KERNEL AND COVARIANCE MATRIX

models dependence explicitly

A positive-semidefinite kernel $k(x, x')$ specifies the covariance between function values. Radial kernels may decay with distance; periodic, linear, and constant kernels need not. Adding independent observation noise contributes only a diagonal term.

$K_{ij} = k(x_i, x_j)$ • $K_y = K + \sigma_n^2 I$ • independence implies diagonal covariance, not necessarily identity

$$\mathbf{f} \mid X, \theta \sim \mathcal{N}(0, K_\theta), \quad \mathbf{y} \mid \mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma_n^2 I) \quad \implies \quad \mathbf{y} \mid X, \theta \sim \mathcal{N}(0, K_y), \quad K_y = K_\theta + \sigma_n^2 I$$

The latent function has been integrated out. Independence implies diagonal noise covariance; identity additionally requires one common unit variance.

The GP marginal likelihood, derived

Just the multivariate Gaussian density, logged

$$\mathbf{y} \mid X, \theta \sim \mathcal{N}(0, K_y(\theta)) : \quad p(\mathbf{y} \mid X, \theta) = (2\pi)^{-N/2} |K_y|^{-1/2} \exp\left(-\frac{1}{2} \mathbf{y}^\top K_y^{-1} \mathbf{y}\right)$$

Step 1: after marginalizing the latent function, one joint Gaussian density replaces an independence product.

$$-\log p(\mathbf{y} \mid X, \theta) = \frac{1}{2} \mathbf{y}^\top K_y^{-1} \mathbf{y} + \frac{1}{2} \log |K_y| + \frac{N}{2} \log 2\pi$$

Step 2: data fit, complexity, constant. Because $\log |K_y|$ depends on θ , it stays.

VERIFICATION

Compare with a library GP value on the same data. A constant offset exposes a mishandled 2π term.

WHEN THE SIMPLE PICTURE BREAKS

Rasmussen and Williams: marginal likelihood

Observations that are not independent

$$\log p(\mathbf{y} \mid X, \theta) = -\frac{1}{2} \mathbf{y}^\top K_y^{-1} \mathbf{y} - \frac{1}{2} \log |K_y| - \frac{N}{2} \log 2\pi$$

Written entirely in K_y , which was defined on the previous slide.

WHY THIS DECOMPOSITION MATTERS

The first term rewards data fit and the second penalizes covariance volume. Their sum supports hyperparameter comparison within the chosen GP family, but does not certify that family or prevent overfitting when many hyperparameters are optimized.

See: [6] Rasmussen & Williams (2006)

WHEN THE SIMPLE PICTURE BREAKS

One GP case, four statistical objects

Classify by what varies, normalizes, and is integrated

$$q_{\theta}(y) := p(y \mid X, \theta) = \int p(y \mid f, \theta) p(f \mid X, \theta) df,$$

$$A : \theta \mapsto p(y \mid f, \theta), \quad C : f \mapsto p(f \mid y, X, \theta),$$

$$B : y \mapsto q_{\theta}(y), \quad D : \theta \mapsto q_{\theta}(y)$$

5-MINUTE CLASSIFICATION ARTIFACT

- ▶ For A through D, record varying and fixed arguments plus the normalization domain
- ▶ Assign likelihood, marginal data density, posterior, and marginal likelihood; then swap and challenge one claim

WHEN THE SIMPLE PICTURE BREAKS

What survives when the library changes

One chain, read in both directions

- ▶ A model can determine an NLL; an objective rarely identifies one model.
- ▶ Likelihood is a function of the parameter, not a distribution over it.
- ▶ Independence has to be earned: the factorisation is a claim about the data.
- ▶ Audit reverse correspondences: which interpretations fit, and what selects among them?

CARRIED INTO TRACKS B AND D

σ stays the physical parameter; the noise scale is σ_n ; λ is the regularization weight, which is where B1's Tikhonov term came from.

PART 05

References

