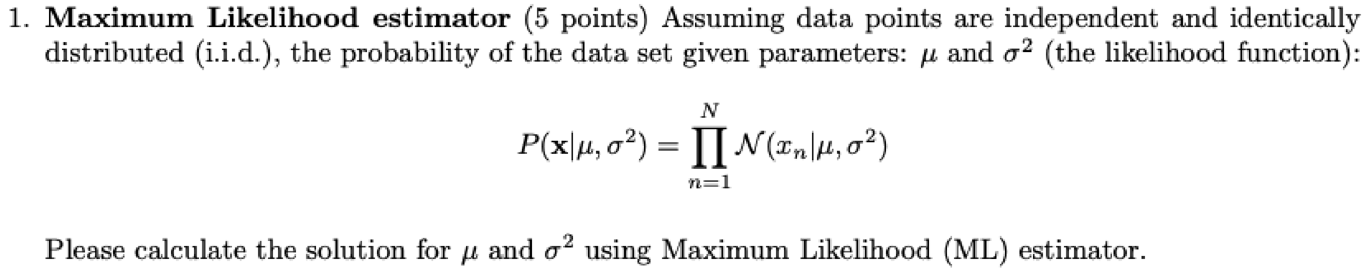
Assignment 2 MLE and MAP (total 20 pts)

Student Name: **Komal Wavhal** (Master in Computer Science)

CWID: **20034443**



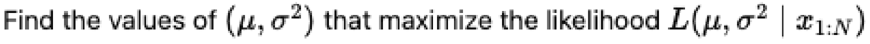
1. **Maximum Likelihood estimator** (5 points) Assuming data points are independent and identically distributed (i.i.d.), the probability of the data set given parameters: μ and σ^2 (the likelihood function):

$$P(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)$$

Please calculate the solution for μ and σ^2 using Maximum Likelihood (ML) estimator.

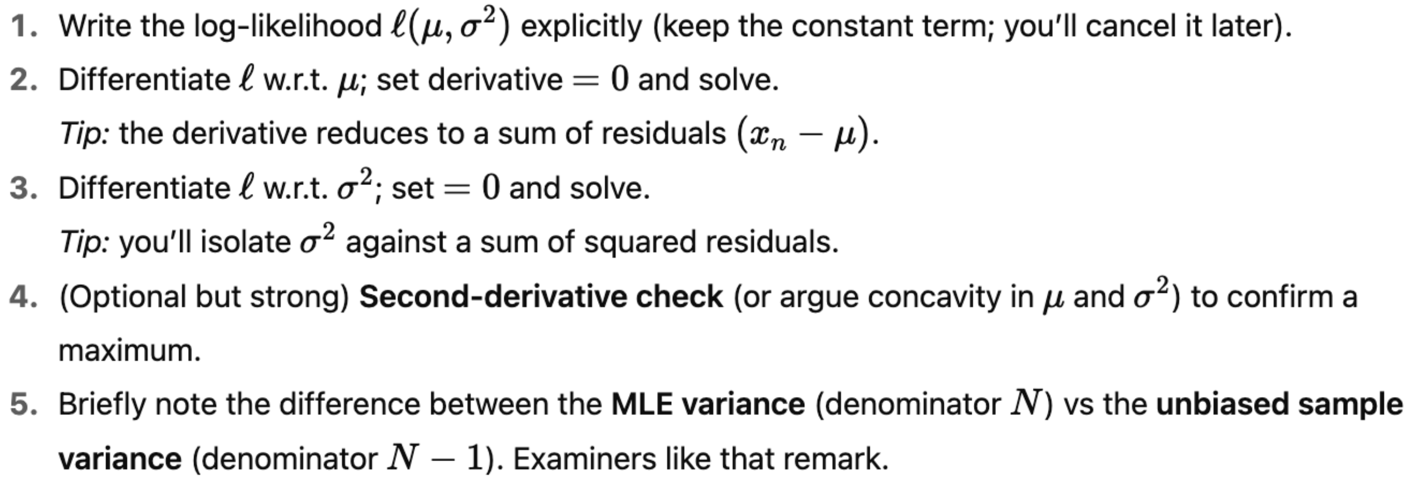
Solution 1:

Goal: Derive closed-form estimators for the mean and variance from the Gaussian likelihood.



Find the values of (μ, σ^2) that maximize the likelihood $L(\mu, \sigma^2 \mid x_{1:N})$

Steps-1: Set up the likelihood and log-likelihood and followed the below steps:



1. Write the log-likelihood $\ell(\mu, \sigma^2)$ explicitly (keep the constant term; you'll cancel it later).
2. Differentiate ℓ w.r.t. μ; set derivative $= 0$ and solve.
   Tip: the derivative reduces to a sum of residuals $(x_n - \mu)$.
3. Differentiate ℓ w.r.t. σ^2; set $= 0$ and solve.
   Tip: you'll isolate σ^2 against a sum of squared residuals.
4. (Optional but strong) **Second-derivative check** (or argue concavity in μ and σ^2) to confirm a maximum.
5. Briefly note the difference between the **MLE variance** (denominator N) vs the **unbiased sample variance** (denominator $N - 1$). Examiners like that remark.

Course 559 A-2 DATE

Solution

Q. 1 For a normal density

$$\mathcal{N}(x \mid \mu\, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{x-\mu}{2\sigma^2}\right)^2$$

Assumming independence:

$$L(\mu, \sigma^2 \mid x_{1:N}) = \prod_{n=1}^{N} \mathcal{N}(x_n \mid \mu, \sigma^2)$$

work with the Log-Likelihood $l = \log L$

$$l(\mu, \sigma^2) = \sum_{n=1}^{N} \log \mathcal{N}(x_n \mid \mu, \sigma^2)$$

$$= \sum_{n=1}^{N} \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{(x_n-\mu)^2}{2\sigma^2} \right]$$

$$= -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n-\mu)^2$$

let $S(\mu) = \sum_{n=1}^{N} (x_n-\mu)^2$ to keep notation Compact

Differenciate w.r.t μ, Set to 0, Solve

$$\frac{\partial l}{\partial \mu} = -\frac{1}{2\sigma^2} \cdot 2 \sum_{n=1}^{N} (x_n-\mu)(-1)$$

$$= \frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n-\mu)$$

Set to zero

$$\sum_{n=1}^{N} (x_n-\mu) = 0 \qquad \hat{\mu}_{ML} = \bar{x} := \frac{1}{N} \sum_{n=1}^{N} x_n$$

2nd derivative in (μ)

$$\frac{\partial^2 l}{\partial \mu^2} = \frac{1}{\sigma^2} \sum_{n=1}^{N} (-1) = -\frac{N}{\sigma^2} < 0$$

So, $\bar{x}$ gives maximum in μ

PAGE

Differenciate w.r.to σ^2

Set to 0, Solve the equation

Treat $v = \sigma^2$ as the Variable using the expression above.

$$l(\mu, v) = -\frac{N}{2}\log(2\pi v) - \frac{1}{2v}S(\mu)$$

Derivative

$$\frac{\partial l}{\partial v} = -\frac{N}{2}\cdot\frac{1}{v} + \frac{1}{2}S(\mu)\cdot\left(-\frac{1}{v^2}\right)$$

with a minus from $\frac{1}{v} = -\frac{N}{2v} + \frac{S(\mu)}{2v^2}$

Set to zero and multiply by $2v^2$

$$S(\mu) - Nv = 0$$

$$\Longrightarrow \hat{\sigma}^2_{ML} = \frac{1}{N}S(\mu)$$

at the MLE, we plug $\mu = \hat{\mu} = \bar{x}$, so

$$\hat{\sigma}^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2$$

2nd derivative Check in ($v = \sigma^2$):

at the Stationary point $v = \hat{v} = S(\mu)/N$,

$$\left.\frac{\partial^2 l}{\partial v^2}\right|_{\hat{v}} = \frac{N}{2}\frac{N^2}{S(\mu)^2} - \frac{S(\mu)}{(S(\mu)/N)^3} = \frac{N^3}{2S(\mu)^2} - \frac{N^3}{S(\mu)^2}$$

$$= -\frac{N^3}{2S(\mu)^2} < 0$$

So, it's a maximum in σ^2 as well.

$$\hat{\mu}_{ML} = \bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n ,$$

$$\hat{\sigma}^2_{ML} = \frac{1}{N}\sum_{n=1}^{N} (x_n - \bar{x})^2$$

MLE variance vs. unbiased sample variance

$$E[\hat{\sigma}^2_{ML}] = \frac{N-1}{N}\sigma^2$$

its a slightly biased low

The unbiased sample variance replases 'N' by N-1

$s^2 := \frac{1}{N-1}\sum_{n=1}^{N} (x_n - \bar{x})^2$ and $E(s^2) = \sigma^2$

Sketch of the expectation result - Using the identity

$$\sum_{n=1}^{N} (x_n - \mu)^2 = \sum_{n=1}^{N} (x_n - \mu)^2 + N(\bar{x} - \mu)^2$$

take expectations under $x_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$

recall $E\left[\sum_{n=1}^{N} (x_n - \mu)^2\right] = N\sigma^2$ and

$$E[(\bar{x} - \mu)^2] = \sigma^2 / N$$

This gives $E[\sum(x_n - \bar{x})^2] = (N-1)\cdot\sigma^2$

Let $x_1, \dots x_N$ be i.i.d $\mathcal{N}(\mu, \sigma^2)$ Thelikelihood and log-likelihood are

$$L(\mu, \sigma^2) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_n-\mu)^2}{2\sigma^2}\right),$$

$$l(\mu, \sigma^2) = \log L = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n-\mu)^2$$

Estimate of μ

$$\frac{\partial l}{\partial \mu} = \frac{1}{2\sigma^2} \cdot 2 \sum_{n=1}^{N} (\mu - x_n) = \frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu)$$

Set to zero $\sum_{n=1}^{N} (x_n - \mu) = 0 \Rightarrow \boxed{\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x_n}$

Second derivative at the optimum:

$$\frac{\partial^2 l}{\partial(\sigma^2)^2} = \frac{N}{(2\sigma^2)^2} = \frac{S(\mu)}{(\sigma^2)^3} \quad s = N\sigma^2$$

$$= -\frac{N}{2(\sigma^2)^2} < 0$$

So, its maximum

This is MLE variance with denominator N the unibiased Variance uses N-1

$$\hat{\mu}_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n \quad \hat{\sigma}_{ML} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \hat{\mu}_{ML})^2$$

2. **Maximum Likelihood** (5 points) We assume there is a true function $f(\mathbf{x})$ and the target value is given by $y = f(x) + \epsilon$ where ϵ is a Gaussian distribution with mean 0 and variance σ^2. Thus,

$$p(y|x, w, \beta) = \mathcal{N}(y|f(x), \beta^{-1})$$

where $\beta^{-1} = \sigma^2$.

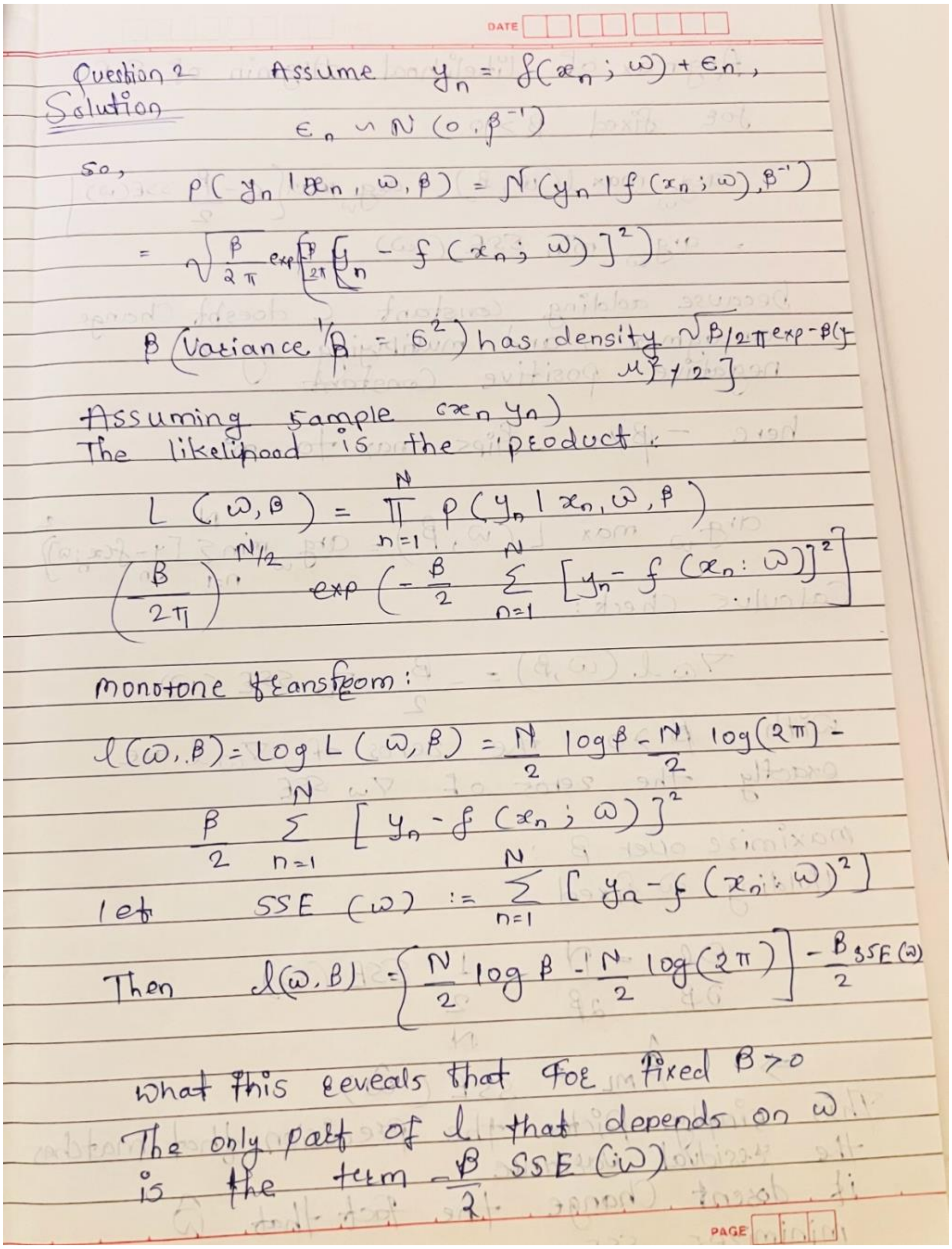
Assuming the data points are drawn independently from the distribution, we obtain the likelihood function:

$$p(\mathbf{y}|\mathbf{x}, w, \beta) = \prod_{n=1}^{N} \mathcal{N}(y_n|f(x), \beta^{-1})$$

Please show that maximizing the likelihood function is equivalent to minimizing the sum-of-squares error function.

Solution 2:

Goal: Prove that maximizing the likelihood under $y_n = f(x_n) + \varepsilon_n$, $\varepsilon_n \sim \mathcal{N}(0, \beta^{-1})$, is equivalent to minimizing the **sum of squared errors (SSE)**.



DATE

Question 2 Solution

Assume $y_n = f(x_n; \omega) + \epsilon_n$,

$\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$

So,

$p(y_n | x_n, \omega, \beta) = \mathcal{N}(y_n | f(x_n; \omega), \beta^{-1})$

$= \sqrt{\frac{\beta}{2\pi}} \exp\left(\frac{\beta}{2\pi}\left[y_n - f(x_n; \omega)\right]^2\right)$

β (Variance $1/\beta = \sigma^2$) has density $\sqrt{\beta/2\pi} \exp - \beta(y - \mu)^2/2]$

Assuming sample $(x_n\, y_n)$
The likelihood is the product:

$L(\omega, \beta) = \prod_{n=1}^{N} p(y_n | x_n, \omega, \beta)$

$\left(\frac{\beta}{2\pi}\right)^{N/2} \cdot \exp\left(-\frac{\beta}{2} \sum_{n=1}^{N} [y_n - f(x_n : \omega)]^2\right]$

monotone transform:

$l(\omega, \beta) = \log L(\omega, \beta) = \frac{N}{2} \log \beta - \frac{N}{2} \log(2\pi) -$

$\frac{\beta}{2} \sum_{n=1}^{N} [y_n - f(x_n; \omega)]^2$

let $SSE(\omega) := \sum_{n=1}^{N} [y_n - f(x_n; \omega)^2]$

Then $l(\omega, \beta) = \left[\frac{N}{2} \log \beta - \frac{N}{2} \log(2\pi)\right] - \frac{\beta}{2} SSE(\omega)$

what this reveals that For fixed $\beta > 0$
The only part of l that depends on ω
is the term $-\frac{\beta}{2} SSE(\omega)$

PAGE

Argmax of likelihood Argmin of SSE

for fixed $\beta > 0$

$$\arg_{\omega} \max l(\omega, \beta) = \arg_{\omega} \max \left\{ C - \frac{\beta}{2} SSE(\omega) \right]$$

$$= \arg_{\omega} \min SSE(\omega)$$

because adding constant C doesnt Change an argmax, and multipiying by a negative positive Constant

here $-\beta/2$ flips max to min.

$$\arg_{\omega} \max L(\omega, \beta) = \arg_{\omega} \min \sum_{n=1}^{N} [y_n - f(x_n; \omega)]^2$$

Calculus Check:

$$\nabla_{\omega} l (\omega, \beta) = -\frac{\beta}{2} \nabla_{\omega} SSE(\omega)$$

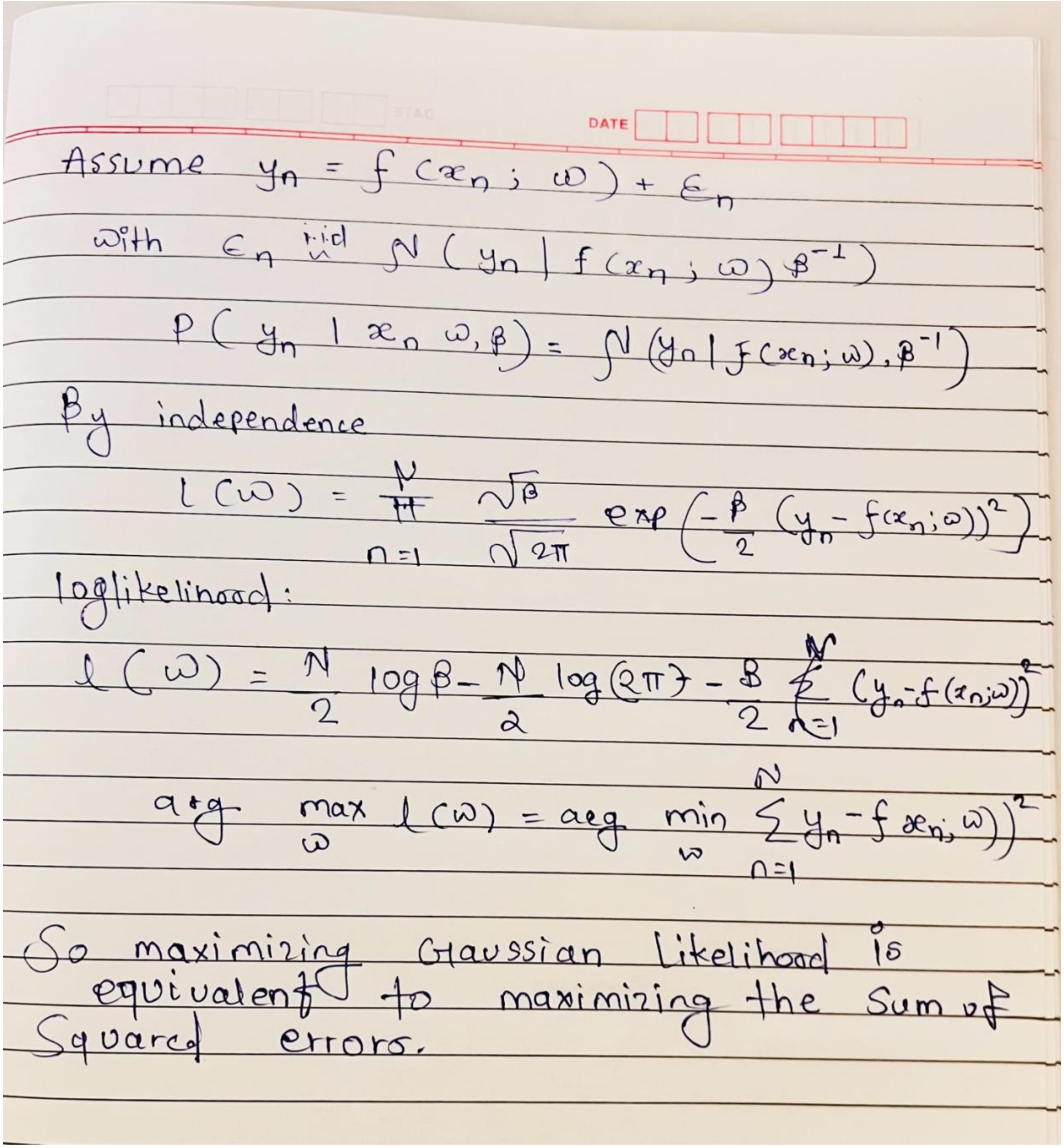
with $\beta > 0$ the zeros of $\nabla_{\omega} l$ are exactly the zeros of $\nabla_{\omega} SSE$

maximize over β:
Holding ω fixed

$$\frac{\partial l}{\partial \beta} = \frac{N}{2\beta} - \frac{1}{2} SSE(\omega)$$

$$\Rightarrow \hat{\beta}_{ML} = \frac{N}{SSE(\hat{\omega})}$$

This simply picks the precision that matches the residual variance
it dosent Change the fact that $\hat{\omega}$ minimizes SSE.



Assume $y_n = f(x_n; \omega) + \epsilon_n$

with $\epsilon_n \overset{i.i.d}{\sim} \mathcal{N}(y_n \mid f(x_n; \omega) \beta^{-1})$

$$P(y_n \mid x_n\, \omega, \beta) = \mathcal{N}(y_n \mid f(x_n; \omega), \beta^{-1})$$

By independence

$$L(\omega) = \prod_{n=1}^{N} \frac{\sqrt{\beta}}{\sqrt{2\pi}} \exp\left(-\frac{\beta}{2}(y_n - f(x_n;\omega))^2\right)$$

loglikelihood:

$$\ell(\omega) = \frac{N}{2}\log\beta - \frac{N}{2}\log(2\pi) - \frac{\beta}{2}\sum_{n=1}^{N}(y_n - f(x_n;\omega))^2$$

$$\text{arg}\max_{\omega} \ell(\omega) = \text{aeg}\min_{w} \sum_{n=1}^{N} y_n - f\, x_n, \omega))^2$$

So maximizing Gaussian Likelihood is equivalent to maximizing the Sum of Squared errors.

3. **MAP estimator** (5 points) Given input values $\mathbf{x} = (x_1, ..., x_N)^T$ and their corresponding target values $\mathbf{y} = (y_1, ..., y_N)^T$, we estimate the target by using function $f(x, \mathbf{w})$ which is a polynomial curve. Assuming the target variables are drawn from Gaussian distribution:

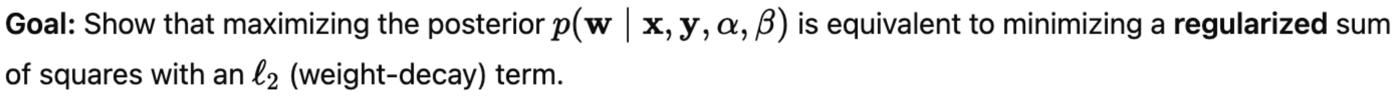
$$p(y|x, \mathbf{w}, \beta) = \mathcal{N}(y|f(x, \mathbf{w}), \beta^{-1})$$

and a prior Gaussian distribution for $\mathbf{w}$:

$$p(\mathbf{w}|\alpha) = (\frac{\alpha}{2\pi})^{(M+1)/2} \exp(-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w})$$

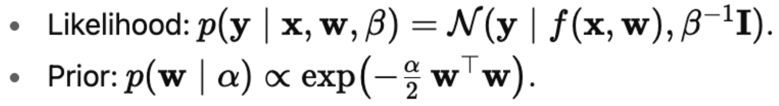
Please prove that maximum posterior (MAP) is equivalent to minimizing the regularized sum-of-squares error function. Note that the posterior distribution of $\mathbf{w}$ is $p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \alpha, \beta)$. **Hint: use Bayes' theorem.**

Solution 3:



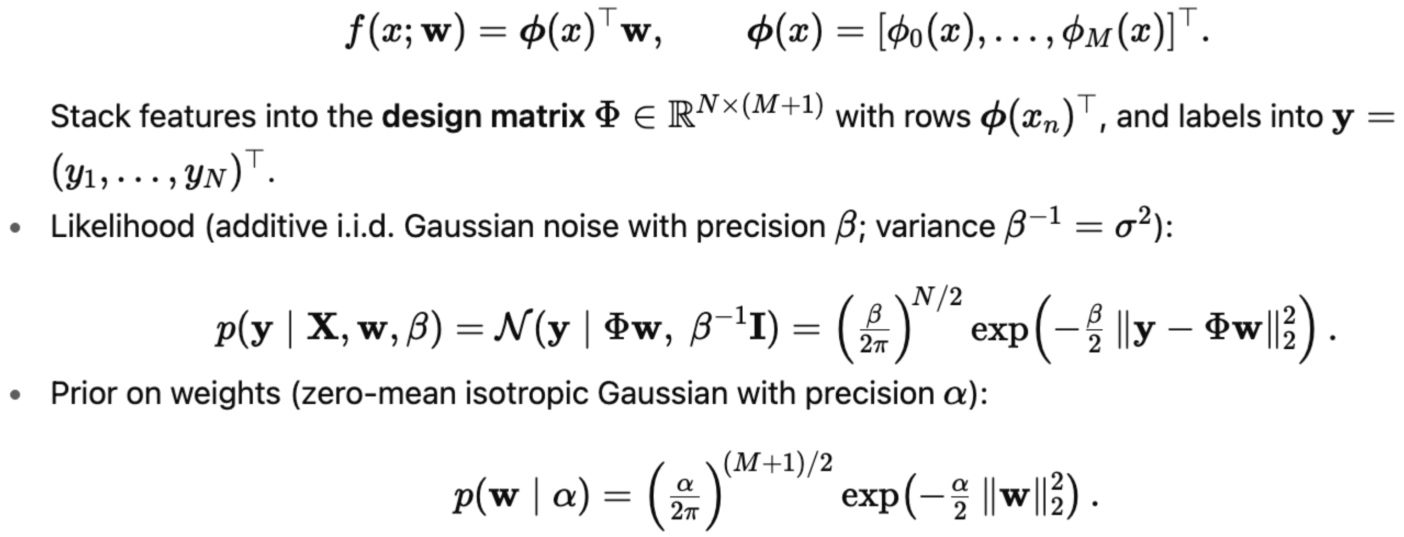
Goal: Show that maximizing the posterior $p(\mathbf{w} \mid \mathbf{x}, \mathbf{y}, \alpha, \beta)$ is equivalent to minimizing a **regularized** sum of squares with an ℓ_2 (weight-decay) term.

Given:



- Likelihood: $p(\mathbf{y} \mid \mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(\mathbf{y} \mid f(\mathbf{x}, \mathbf{w}), \beta^{-1}\mathbf{I})$.
- Prior: $p(\mathbf{w} \mid \alpha) \propto \exp\left(-\frac{\alpha}{2}\, \mathbf{w}^\top \mathbf{w}\right)$.

Model (polynomial curve, but any linear-in-parameters model works):



$$f(x; \mathbf{w}) = \boldsymbol{\phi}(x)^\top \mathbf{w}, \qquad \boldsymbol{\phi}(x) = [\phi_0(x), \ldots, \phi_M(x)]^\top.$$

Stack features into the **design matrix** $\boldsymbol{\Phi} \in \mathbb{R}^{N \times (M+1)}$ with rows $\boldsymbol{\phi}(x_n)^\top$, and labels into $\mathbf{y} = (y_1, \ldots, y_N)^\top$.

- Likelihood (additive i.i.d. Gaussian noise with precision β; variance $\beta^{-1} = \sigma^2$):

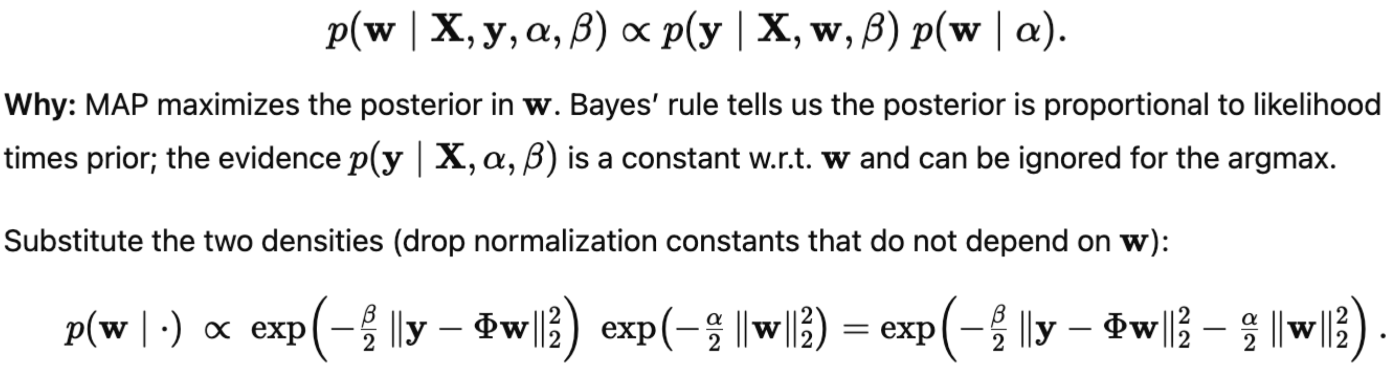
$$p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \beta) = \mathcal{N}(\mathbf{y} \mid \boldsymbol{\Phi}\mathbf{w},\ \beta^{-1}\mathbf{I}) = \left(\frac{\beta}{2\pi}\right)^{N/2} \exp\left(-\frac{\beta}{2}\, \|\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}\|_2^2\right).$$

- Prior on weights (zero-mean isotropic Gaussian with precision α):

$$p(\mathbf{w} \mid \alpha) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left(-\frac{\alpha}{2}\, \|\mathbf{w}\|_2^2\right).$$

Using precisions α, β keeps constants clean and makes the posterior exponent a quadratic in w, which ensures a Gaussian posterior and a convex optimization.

Step 1 — Bayes' rule to write the posterior



$$p(\mathbf{w} \mid \mathbf{X}, \mathbf{y}, \alpha, \beta) \propto p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \beta)\; p(\mathbf{w} \mid \alpha).$$

Why: MAP maximizes the posterior in $\mathbf{w}$. Bayes' rule tells us the posterior is proportional to likelihood times prior; the evidence $p(\mathbf{y} \mid \mathbf{X}, \alpha, \beta)$ is a constant w.r.t. $\mathbf{w}$ and can be ignored for the argmax.

Substitute the two densities (drop normalization constants that do not depend on $\mathbf{w}$):

$$p(\mathbf{w} \mid \cdot)\ \propto\ \exp\left(-\frac{\beta}{2}\, \|\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}\|_2^2\right)\ \exp\left(-\frac{\alpha}{2}\, \|\mathbf{w}\|_2^2\right) = \exp\left(-\frac{\beta}{2}\, \|\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}\|_2^2 - \frac{\alpha}{2}\, \|\mathbf{w}\|_2^2\right).$$

Step 2 — Take negative log posterior

Because $\log(\cdot)$ is strictly increasing, maximizing the posterior equals minimizing the **negative log-posterior**:

$$\mathcal{L}(\mathbf{w}) := -\log p(\mathbf{w} \mid \mathbf{X}, \mathbf{y}, \alpha, \beta) = \tfrac{\beta}{2}\,\|\mathbf{y} - \mathbf{\Phi}\mathbf{w}\|_2^2 + \tfrac{\alpha}{2}\,\|\mathbf{w}\|_2^2 + \text{const.}$$

Step 3 — Identify the regularized least-squares objective

Divide by $\beta/2 > 0$ (a positive scalar doesn't change the minimizer) to obtain:

$$\boxed{J(\mathbf{w}) = \|\mathbf{y} - \mathbf{\Phi}\mathbf{w}\|_2^2 \;+\; \lambda\,\|\mathbf{w}\|_2^2, \qquad \lambda = \frac{\alpha}{\beta}}.$$

This is exactly **ridge regression** (squared-error loss with ℓ_2 weight decay).

Therefore,

$$\boxed{\mathbf{w}_{\text{MAP}} = \arg\max_{\mathbf{w}} p(\mathbf{w} \mid \mathbf{X}, \mathbf{y}, \alpha, \beta) = \arg\min_{\mathbf{w}}\; \|\mathbf{y} - \mathbf{\Phi}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2}.$$

This interprets that:

- The likelihood term pushes Φw to fit the data (small squared residuals).
- The Gaussian prior shrinks weights toward zero; α controls shrinkage strength.
- λ=α/β balances fit vs. complexity: larger α (tighter prior) or smaller β (noisier data) → stronger regularization.

Step 4 — Closed-form MAP solution

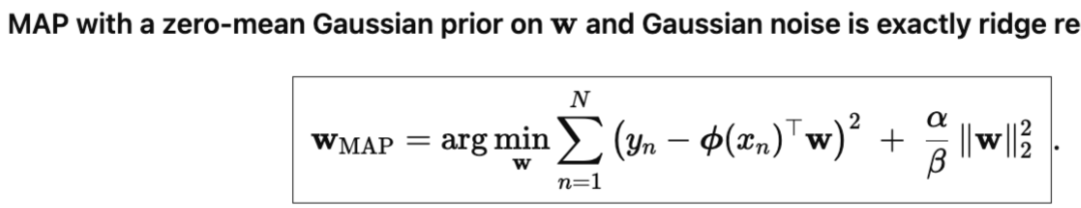
Since $J(\mathbf{w})$ is a strictly convex quadratic, the unique minimizer satisfies the normal equations:

$$(\mathbf{\Phi}^\top\mathbf{\Phi} + \lambda\mathbf{I})\,\mathbf{w} = \mathbf{\Phi}^\top\mathbf{y} \quad \Rightarrow \quad \boxed{\mathbf{w}_{\text{MAP}} = (\mathbf{\Phi}^\top\mathbf{\Phi} + \lambda\mathbf{I})^{-1}\mathbf{\Phi}^\top\mathbf{y}}.$$

Equivalently, in precision form:

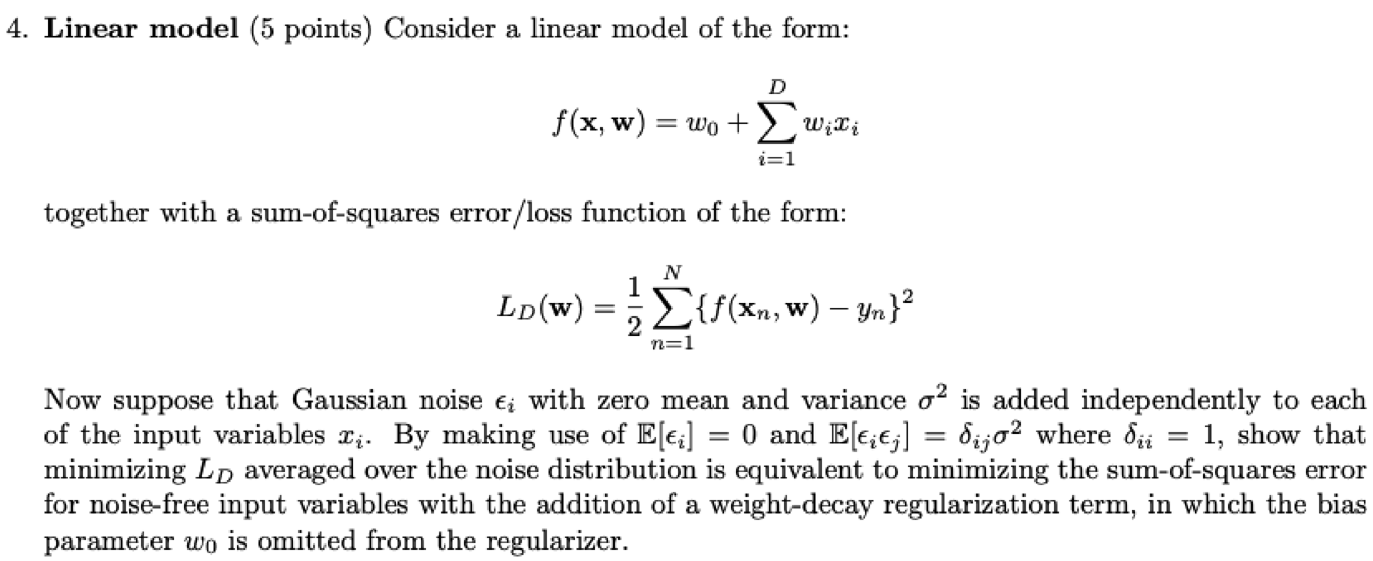
$$\mathbf{w}_{\text{MAP}} = (\alpha\mathbf{I} + \beta\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\,\beta\mathbf{\Phi}^\top\mathbf{y}.$$

Why this form: It's the minimizer of a positive-definite quadratic; the Hessian is $\nabla^2 J = 2(\mathbf{\Phi}^\top\mathbf{\Phi} + \lambda\mathbf{I}) \succ 0$.



MAP with a zero-mean Gaussian prior on $\mathbf{w}$ and Gaussian noise is exactly ridge regression:

$$\boxed{\mathbf{w}_{\text{MAP}} = \arg\min_{\mathbf{w}} \sum_{n=1}^{N} \left(y_n - \boldsymbol{\phi}(x_n)^\top \mathbf{w}\right)^2 \;+\; \frac{\alpha}{\beta}\,\|\mathbf{w}\|_2^2}\,.$$



4. **Linear model** (5 points) Consider a linear model of the form:

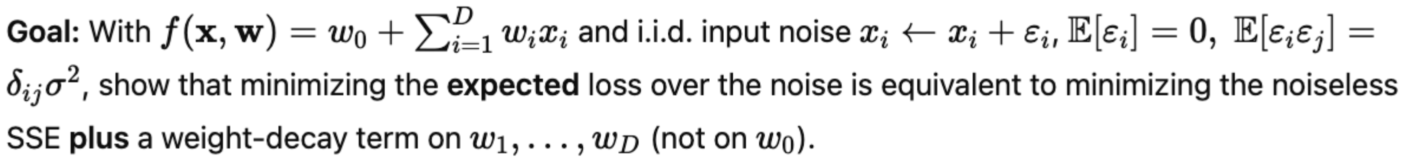
$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i$$

together with a sum-of-squares error/loss function of the form:

$$L_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{f(\mathbf{x}_n, \mathbf{w}) - y_n\}^2$$

Now suppose that Gaussian noise ϵ_i with zero mean and variance σ^2 is added independently to each of the input variables x_i. By making use of $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i \epsilon_j] = \delta_{ij}\sigma^2$ where $\delta_{ii} = 1$, show that minimizing L_D averaged over the noise distribution is equivalent to minimizing the sum-of-squares error for noise-free input variables with the addition of a weight-decay regularization term, in which the bias parameter w_0 is omitted from the regularizer.

Solution 4:



Goal: With $f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i$ and i.i.d. input noise $x_i \leftarrow x_i + \varepsilon_i$, $\mathbb{E}[\varepsilon_i] = 0,\ \mathbb{E}[\varepsilon_i \varepsilon_j] = \delta_{ij}\sigma^2$, show that minimizing the **expected** loss over the noise is equivalent to minimizing the noiseless SSE **plus** a weight-decay term on $w_1, \ldots, w_D$ (not on w_0).

Linear model

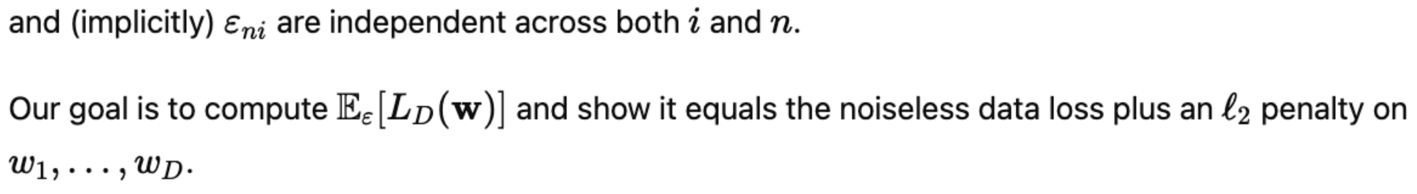
$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i$$

$$L_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(f(\mathbf{x}_n, \mathbf{w}) - y_n\right)^2.$$

with training loss

disturb **each input coordinate** by independent zero-mean Gaussian noise

$$\tilde{x}_{ni} = x_{ni} + \varepsilon_{ni}, \qquad \mathbb{E}[\varepsilon_{ni}] = 0, \quad \mathbb{E}[\varepsilon_{ni}\varepsilon_{nj}] = \delta_{ij}\sigma^2,$$



and (implicitly) ε_{ni} are independent across both i and n.

Our goal is to compute $\mathbb{E}_\varepsilon[L_D(\mathbf{w})]$ and show it equals the noiseless data loss plus an ℓ_2 penalty on $w_1, \ldots, w_D$.

Substitute the noisy inputs

With noise, the prediction on sample n is

$$f(\tilde{\mathbf{x}}_n, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i(x_{ni} + \varepsilon_{ni}) = \underbrace{\Big(w_0 + \sum_{i=1}^{D} w_i x_{ni}\Big)}_{=: a_n} + \sum_{i=1}^{D} w_i \varepsilon_{ni}.$$

Defined a =f(xn,w) as the **noiseless** prediction to separate deterministic and random parts.

Take expectation over the input noise

Since $\mathbb{E}[\varepsilon_{ni}] = 0$,

$$\mathbb{E}\left[2(a_n - y_n)\sum_{i=1}^{D} w_i \varepsilon_{ni}\right] = 2(a_n - y_n)\sum_{i=1}^{D} w_i \underbrace{\mathbb{E}[\varepsilon_{ni}]}_{0} = 0.$$

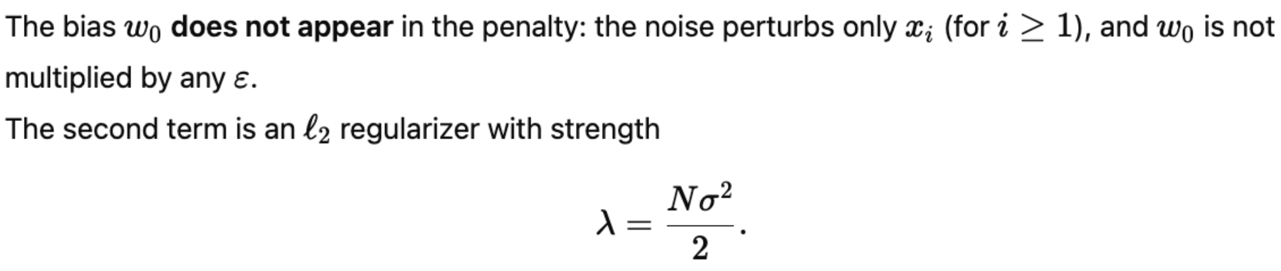
The **cross term** vanishes (deterministic factor times zero-mean noise).

For the quadratic noise term, use independence and the given covariance:

$$\begin{aligned}
\mathbb{E}\left[\Big(\sum_{i=1}^{D} w_i \varepsilon_{ni}\Big)^2\right] &= \mathbb{E}\left[\sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j \varepsilon_{ni}\varepsilon_{nj}\right] \\
&= \sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j \, \mathbb{E}[\varepsilon_{ni}\varepsilon_{nj}] \\
&= \sum_{i=1}^{D} w_i^2 \, \sigma^2 \qquad (\text{since } \mathbb{E}[\varepsilon_{ni}\varepsilon_{nj}] = \delta_{ij}\sigma^2).
\end{aligned}$$

$$\mathbb{E}_\varepsilon\left[\big(f(\tilde{\mathbf{x}}_n, \mathbf{w}) - y_n\big)^2\right] = (a_n - y_n)^2 + \sigma^2 \sum_{i=1}^{D} w_i^2.$$

$$\begin{aligned}
\mathbb{E}_\varepsilon[L_D(\mathbf{w})] &= \frac{1}{2}\sum_{n=1}^{N} \mathbb{E}_\varepsilon\left[\big(f(\tilde{\mathbf{x}}_n, \mathbf{w}) - y_n\big)^2\right] \\
&= \frac{1}{2}\sum_{n=1}^{N}(a_n - y_n)^2 + \frac{1}{2}\sum_{n=1}^{N}\sigma^2 \sum_{i=1}^{D} w_i^2 \\
&= \underbrace{\frac{1}{2}\sum_{n=1}^{N}\big(f(\mathbf{x}_n, \mathbf{w}) - y_n\big)^2}_{\text{noiseless SSE}} + \underbrace{\frac{N\sigma^2}{2}\sum_{i=1}^{D} w_i^2}_{\text{weight decay on } w_1, \ldots, w_D} .
\end{aligned}$$

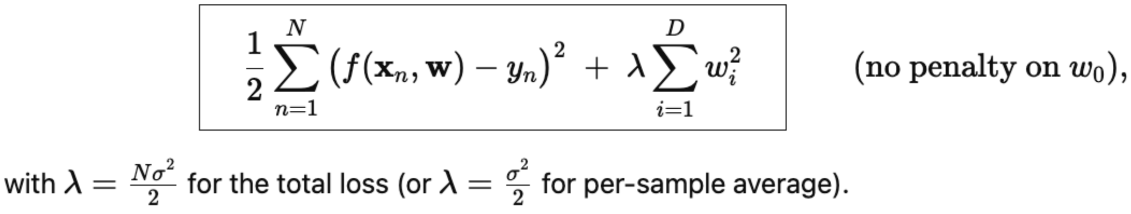


The bias w_0 **does not appear** in the penalty: the noise perturbs only x_i (for $i \geq 1$), and w_0 is not multiplied by any ε.

The second term is an ℓ_2 regularizer with strength

$$\lambda = \frac{N\sigma^2}{2}.$$

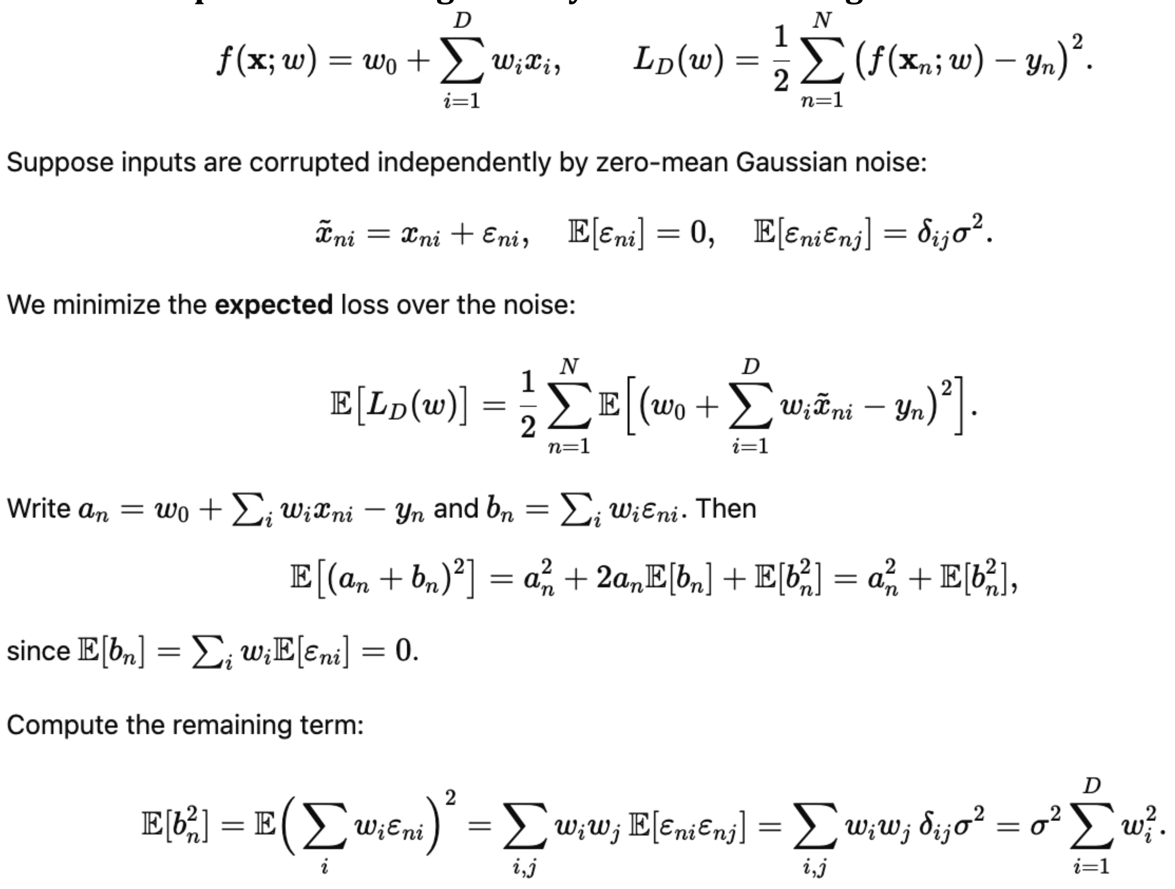
Minimizing the expected noisy-input loss is **exactly equivalent** to minimizing, on the **noise-free** inputs, the regularized objective



$$\boxed{\frac{1}{2}\sum_{n=1}^{N}\big(f(\mathbf{x}_n,\mathbf{w})-y_n\big)^2 \;+\; \lambda\sum_{i=1}^{D} w_i^2} \qquad \text{(no penalty on } w_0\text{)},$$

with $\lambda = \frac{N\sigma^2}{2}$ for the total loss (or $\lambda = \frac{\sigma^2}{2}$ for per-sample average).

Linear model with input noise □ weight decay on non-bias weights



$$f(\mathbf{x}; w) = w_0 + \sum_{i=1}^{D} w_i x_i, \qquad L_D(w) = \frac{1}{2}\sum_{n=1}^{N}\big(f(\mathbf{x}_n; w) - y_n\big)^2.$$

Suppose inputs are corrupted independently by zero-mean Gaussian noise:

$$\tilde{x}_{ni} = x_{ni} + \varepsilon_{ni}, \quad \mathbb{E}[\varepsilon_{ni}] = 0, \quad \mathbb{E}[\varepsilon_{ni}\varepsilon_{nj}] = \delta_{ij}\sigma^2.$$

We minimize the **expected** loss over the noise:

$$\mathbb{E}\big[L_D(w)\big] = \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}\Big[\Big(w_0 + \sum_{i=1}^{D} w_i\tilde{x}_{ni} - y_n\Big)^2\Big].$$

Write $a_n = w_0 + \sum_i w_i x_{ni} - y_n$ and $b_n = \sum_i w_i \varepsilon_{ni}$. Then

$$\mathbb{E}\big[(a_n + b_n)^2\big] = a_n^2 + 2a_n\mathbb{E}[b_n] + \mathbb{E}[b_n^2] = a_n^2 + \mathbb{E}[b_n^2],$$

since $\mathbb{E}[b_n] = \sum_i w_i \mathbb{E}[\varepsilon_{ni}] = 0$.

Compute the remaining term:

$$\mathbb{E}[b_n^2] = \mathbb{E}\Big(\sum_i w_i\varepsilon_{ni}\Big)^2 = \sum_{i,j} w_i w_j\,\mathbb{E}[\varepsilon_{ni}\varepsilon_{nj}] = \sum_{i,j} w_i w_j\,\delta_{ij}\sigma^2 = \sigma^2\sum_{i=1}^{D} w_i^2.$$

Only the weights attached to **noisy inputs** (w1,…,wD) are penalized; the **bias w0 is not** (it doesn't multiply the noise).

Equivalently, minimizing expected noisy loss is the same (up to a constant factor) as minimizing

$$\mathbb{E}\big[L_D(w)\big] = \frac{1}{2}\sum_{n=1}^{N} a_n^2 + \frac{1}{2}\sum_{n=1}^{N}\sigma^2\sum_{i=1}^{D} w_i^2 = \underbrace{\frac{1}{2}\sum_{n=1}^{N}\Big(w_0 + \sum_i w_i x_{ni} - y_n\Big)^2}_{\text{noiseless SSE}} + \underbrace{\frac{N\sigma^2}{2}\sum_{i=1}^{D} w_i^2}_{\text{weight decay}}.$$

Therefore:

$$\boxed{\frac{1}{2}\sum_{n=1}^{N}\Big(w_0 + \sum_i w_i x_{ni} - y_n\Big)^2 + \lambda\sum_{i=1}^{D} w_i^2, \quad \lambda = \frac{N\sigma^2}{2}\ (> 0).}$$