Chapter 1

Linear Regression

Ordinary Least Squares

Theorem 1.0.1 (Uniqueness of the Least Squares Solution). *Let $\Phi \in \mathbb{R}^{N\times M}$ denote the design matrix and $t \in \mathbb{R}^N$ the target vector. Consider the least squares cost function*

$$E(w) = \frac{1}{2}\|t - \Phi w\|^2.$$

Then:

(i) The function $E(w)$ is convex in w.

(ii) If $\Phi^\top \Phi$ is invertible (i.e., $\operatorname{rank}(\Phi) = M$), then $E(w)$ is strictly convex *and admits a unique minimizer*

$$w^* = (\Phi^\top \Phi)^{-1}\Phi^\top t.$$

(iii) If $\Phi^\top\Phi$ is singular, the minimizer is not unique; all minimizers are of the form

$$w = w_0 + v, \qquad v \in \operatorname{Null}(\Phi),$$

where w_0 is any particular solution to the normal equations $\Phi^\top \Phi w = \Phi^\top t$.

Proof. We begin by expanding the objective:

$$E(w) = \frac{1}{2}(t - \Phi w)^\top (t - \Phi w) = \frac{1}{2}(t^\top t - 2t^\top \Phi w + w^\top \Phi^\top \Phi w).$$

(1) Gradient and Stationary Point: The gradient of $E(w)$ with respect to w is

$$\nabla_w E(w) = -\Phi^\top t + \Phi^\top \Phi w.$$

Setting $\nabla_w E(w) = 0$ yields the *normal equations*

$$\Phi^\top \Phi w = \Phi^\top t. \tag{1}$$

(2) Hessian and Convexity: The Hessian of $E(w)$ is

$$H = \nabla_w^2 E(w) = \Phi^\top \Phi.$$

For any nonzero vector $z \in \mathbb{R}^M$,

$$z^\top H z = z^\top \Phi^\top \Phi z = \|\Phi z\|^2 \geq 0,$$

hence H is positive semidefinite, implying $E(w)$ is convex.
If Φ has full column rank ($\operatorname{rank}(\Phi) = M$), then $\Phi^\top\Phi$ is positive definite, and

$$z^\top H z = 0 \quad \Leftrightarrow \quad z = 0,$$

so $E(w)$ is strictly convex. A strictly convex function has a unique minimizer, obtained by solving (1):

$$w^* = (\Phi^\top \Phi)^{-1}\Phi^\top t.$$

(3) Non-uniqueness for Rank-Deficient Φ: If $\Phi^\top\Phi$ is singular, there exist nonzero vectors v such that $\Phi v = 0$. For any particular solution w_0 satisfying (1), we have

$$\Phi^\top \Phi(w_0 + v) = \Phi^\top \Phi w_0 + \Phi^\top \Phi v = \Phi^\top t,$$

since $\Phi v = 0$. Thus, every vector $w = w_0 + v$, with $v \in \operatorname{Null}(\Phi)$, minimizes $E(w)$. The minimal-norm solution among them is given by the Moore–Penrose pseudoinverse:

$$w^* = \Phi^+ t.$$

(4) Conclusion: The cost $E(w)$ is convex for all Φ, and strictly convex (hence uniquely minimized) iff $\Phi^\top\Phi$ is invertible. □

Theorem 1.0.2 (Unbiasedness of the OLS Estimator). *Assume the linear regression model*

$$t = \Phi w + \varepsilon,$$

where $\Phi \in \mathbb{R}^{N\times M}$ is the design matrix, $w \in \mathbb{R}^M$ the true parameter vector, and the noise satisfies $\mathbb{E}[\varepsilon] = 0$ and $\operatorname{Cov}(\varepsilon) = \sigma^2 I$. Assume further that $\Phi^\top\Phi$ is invertible. Then the ordinary least squares estimator

$$\hat{w} = (\Phi^\top \Phi)^{-1}\Phi^\top t$$

is an unbiased estimator of w, i.e.

$$\mathbb{E}[\hat{w}] = w.$$

Proof. By the model,

$$t = \Phi w + \varepsilon.$$

Substitute into the estimator:

$$\hat{w} = (\Phi^\top \Phi)^{-1}\Phi^\top t = (\Phi^\top \Phi)^{-1}\Phi^\top (\Phi w + \varepsilon).$$

Distribute terms:

$$\hat{w} = (\Phi^\top \Phi)^{-1}\Phi^\top \Phi\, w \; + \; (\Phi^\top \Phi)^{-1}\Phi^\top \varepsilon.$$

Since $(\Phi^\top \Phi)^{-1}\Phi^\top \Phi = I_M$, this simplifies to

$$\hat{w} = w + (\Phi^\top \Phi)^{-1}\Phi^\top \varepsilon.$$

Take expectation using linearity and $\mathbb{E}[\varepsilon] = 0$:

$$\mathbb{E}[\hat{w}] = \mathbb{E}\big[w + (\Phi^\top \Phi)^{-1}\Phi^\top \varepsilon\big] = w + (\Phi^\top \Phi)^{-1}\Phi^\top \mathbb{E}[\varepsilon] = w + (\Phi^\top \Phi)^{-1}\Phi^\top\, 0 = w.$$

Thus $\hat{w}$ is unbiased. □

Corollary 1.0.3. *Under the same assumptions,*

$$\mathrm{Cov}(\hat{w}) = \sigma^2 (\Phi^\top \Phi)^{-1}.$$

Proof. From $\hat{w} = w + (\Phi^\top \Phi)^{-1}\Phi^\top \varepsilon$ and $\mathrm{Cov}(\varepsilon) = \sigma^2 I$,

$$\mathrm{Cov}(\hat{w}) = (\Phi^\top \Phi)^{-1}\Phi^\top\, \mathrm{Cov}(\varepsilon)\, \Phi(\Phi^\top \Phi)^{-1} = \sigma^2 (\Phi^\top \Phi)^{-1}\Phi^\top \Phi(\Phi^\top \Phi)^{-1} = \sigma^2 (\Phi^\top \Phi)^{-1}.$$

□

Theorem 1.0.4 (Covariance of the OLS Estimator)**.** *Under the linear regression model*

$$t = \Phi w + \varepsilon, \qquad \mathbb{E}[\varepsilon] = 0, \quad \mathrm{Cov}(\varepsilon) = \sigma^2 I,$$

with $\Phi \in \mathbb{R}^{N\times M}$ of full column rank, the ordinary least squares estimator

$$\hat{w} = (\Phi^\top \Phi)^{-1}\Phi^\top t$$

has covariance matrix

$$\mathrm{Cov}(\hat{w}) = \sigma^2 (\Phi^\top \Phi)^{-1}.$$

Proof. From the model $t = \Phi w + \varepsilon$,

$$\hat{w} = (\Phi^\top \Phi)^{-1}\Phi^\top t = (\Phi^\top \Phi)^{-1}\Phi^\top (\Phi w + \varepsilon) = w + (\Phi^\top \Phi)^{-1}\Phi^\top \varepsilon.$$

Subtract the expectation $\mathbb{E}[\hat{w}] = w$ to get the deviation:

$$\hat{w} - \mathbb{E}[\hat{w}] = (\Phi^\top \Phi)^{-1}\Phi^\top \varepsilon.$$

Now compute the covariance:

$$\begin{aligned}\mathrm{Cov}(\hat{w}) &= \mathbb{E}\big[(\hat{w} - \mathbb{E}[\hat{w}])(\hat{w} - \mathbb{E}[\hat{w}])^\top\big] \\ &= \mathbb{E}\big[(\Phi^\top \Phi)^{-1}\Phi^\top \varepsilon\varepsilon^\top \Phi(\Phi^\top \Phi)^{-1}\big].\end{aligned}$$

Using $\mathrm{Cov}(\varepsilon) = \sigma^2 I$ and the linearity of expectation:

$$\mathrm{Cov}(\hat{w}) = (\Phi^\top \Phi)^{-1}\Phi^\top (\sigma^2 I)\Phi(\Phi^\top \Phi)^{-1} = \sigma^2 (\Phi^\top \Phi)^{-1}\Phi^\top \Phi(\Phi^\top \Phi)^{-1}.$$

Simplifying:

$$\boxed{\mathrm{Cov}(\hat{w}) = \sigma^2 (\Phi^\top \Phi)^{-1}}.$$

□

Theorem 1.0.5 (Gauss–Markov Theorem)**.** *Consider the linear model*

$$t = \Phi w + \varepsilon,$$

with $\Phi \in \mathbb{R}^{N\times M}$ of full column rank, $\mathbb{E}[\varepsilon] = 0$, and $\mathrm{Cov}(\varepsilon) = \sigma^2 I$. Let $\hat{w}_{\mathrm{OLS}} = (\Phi^\top \Phi)^{-1}\Phi^\top t$ denote the ordinary least squares estimator. Then $\hat{w}_{\mathrm{OLS}}$ is the Best Linear Unbiased Estimator (BLUE)*: for any other linear unbiased estimator of the form $\tilde{w} = Ct$ (with constant matrix $C \in \mathbb{R}^{M\times N}$ such that $\mathbb{E}[\tilde{w}] = w$), we have*

$$\mathrm{Cov}(\tilde{w}) - \mathrm{Cov}(\hat{w}_{\mathrm{OLS}}) \;\succeq\; 0,$$

i.e. the matrix difference is positive semidefinite. Equivalently, every componentwise variance of $\tilde{w}$ is at least that of $\hat{w}_{\mathrm{OLS}}$.

Proof. Let $\tilde{w}$ be any linear estimator of the form $\tilde{w} = Ct$ for a fixed matrix $C \in \mathbb{R}^{M\times N}$. The unbiasedness condition $\mathbb{E}[\tilde{w}] = w$ requires

$$\mathbb{E}[Ct] = C\mathbb{E}[t] = C\Phi w = w \quad \text{for all } w,$$

hence

$$C\Phi = I_M. \tag{1}$$

Write the OLS estimator as

$$\hat{w} \equiv \hat{w}_{\mathrm{OLS}} = (\Phi^\top \Phi)^{-1}\Phi^\top t.$$

Define the matrix difference

$$A \;:=\; C - (\Phi^\top \Phi)^{-1}\Phi^\top.$$

Using (1) and the identity $\big((\Phi^\top \Phi)^{-1}\Phi^\top\big)\Phi = I_M$, we obtain

$$A\Phi = C\Phi - (\Phi^\top \Phi)^{-1}\Phi^\top \Phi = I_M - I_M = 0.$$

Thus

$$A\Phi = 0 \qquad \Longrightarrow \qquad A\Phi w = 0 \quad \text{for all } w.$$

Now express $\tilde{w}$ in terms of $\hat{w}$ and A:

$$\tilde{w} = Ct = \big((\Phi^\top \Phi)^{-1}\Phi^\top + A\big)t = \hat{w} + At.$$

Subtracting expectations (and using $\mathbb{E}[\hat{w}] = \mathbb{E}[\tilde{w}] = w$) gives the zero-mean deviations

$$\tilde{w} - w = (\hat{w} - w) + A\varepsilon,$$

since $t = \Phi w + \varepsilon$ and $A\Phi w = 0$.
Compute the covariance matrices. Using $\mathrm{Cov}(\varepsilon) = \sigma^2 I$ and independence of deterministic matrices from ε,

$$\begin{aligned}\mathrm{Cov}(\tilde{w}) &= \mathbb{E}\big[(\tilde{w} - w)(\tilde{w} - w)^\top\big] \\ &= \mathbb{E}\big[(\hat{w} - w + A\varepsilon)(\hat{w} - w + A\varepsilon)^\top\big] \\ &= \mathrm{Cov}(\hat{w}) + A\,\mathbb{E}[\varepsilon\varepsilon^\top]\,A^\top \;+\; \mathbb{E}\big[(\hat{w} - w)\varepsilon^\top\big]A^\top \;+\; A\,\mathbb{E}\big[\varepsilon(\hat{w} - w)^\top\big].\end{aligned}$$

But $\hat{w} - w = (\Phi^\top\Phi)^{-1}\Phi^\top\varepsilon$ is linear in ε, so

$$\mathbb{E}\big[(\hat{w} - w)\varepsilon^\top\big] = (\Phi^\top\Phi)^{-1}\Phi^\top\mathbb{E}[\varepsilon\varepsilon^\top] = (\Phi^\top\Phi)^{-1}\Phi^\top(\sigma^2 I) = \sigma^2(\Phi^\top\Phi)^{-1}\Phi^\top.$$

Since $A\Phi = 0$, we have

$$\mathbb{E}\big[(\hat{w} - w)\varepsilon^\top\big]A^\top = \sigma^2(\Phi^\top\Phi)^{-1}\Phi^\top A^\top = \sigma^2(\Phi^\top\Phi)^{-1}(\Phi^\top A^\top) = \sigma^2(\Phi^\top\Phi)^{-1}(A\Phi)^\top = 0.$$

Similarly the other cross term $A\,\mathbb{E}[\varepsilon(\hat{w} - w)^\top]$ vanishes. Thus the covariance simplifies to

$$\mathrm{Cov}(\tilde{w}) = \mathrm{Cov}(\hat{w}) + A\,\mathbb{E}[\varepsilon\varepsilon^\top]\,A^\top = \mathrm{Cov}(\hat{w}) + \sigma^2 AA^\top.$$

Therefore

$$\mathrm{Cov}(\tilde{w}) - \mathrm{Cov}(\hat{w}) = \sigma^2 AA^\top.$$

But $\sigma^2 AA^\top$ is positive semidefinite (for any $\sigma^2 \geq 0$ and any matrix A), so

$$\mathrm{Cov}(\tilde{w}) - \mathrm{Cov}(\hat{w}) \succeq 0,$$

which proves that $\hat{w}$ has the smallest covariance matrix among all linear unbiased estimators. This completes the proof. □

Theorem 1.0.6 (Orthogonality of Residuals). *Let $\Phi \in \mathbb{R}^{N\times M}$ be the design matrix and $t \in \mathbb{R}^N$ the observed targets. Let $\hat{w}$ be any solution of the normal equations*

$$\Phi^\top\Phi\,\hat{w} \;=\; \Phi^\top t.$$

Define the residual vector $r := t - \Phi\hat{w}$. Then

$$\Phi^\top r \;=\; 0,$$

i.e. r is orthogonal to every column of Φ (equivalently r is orthogonal to $\mathrm{col}(\Phi)$).

Proof. Starting from the normal equations,

$$\Phi^\top\Phi\,\hat{w} \;=\; \Phi^\top t.$$

Rearrange terms to move $\Phi^\top\Phi\hat{w}$ to the right-hand side:

$$\Phi^\top t - \Phi^\top\Phi\,\hat{w} \;=\; 0.$$

Factor $\Phi^\top$:

$$\Phi^\top(t - \Phi\hat{w}) \;=\; 0.$$

But $t - \Phi\hat{w}$ is exactly the residual vector r, hence

$$\Phi^\top r \;=\; 0.$$

This shows each column of Φ has zero inner product with r, i.e. $r \perp \mathrm{col}(\Phi)$. □

Corollary 1.0.7 (Hat Matrix and Residual Projection). *If Φ has full column rank and $\hat{w} = (\Phi^\top\Phi)^{-1}\Phi^\top t$, define the hat (projection) matrix*

$$P := \Phi(\Phi^\top\Phi)^{-1}\Phi^\top.$$

Then the fitted values are $\hat{t} = Pt$ and the residual satisfies

$$r = (I - P)t,$$

with $P^2 = P$ and $P^\top = P$. Consequently $(I - P)$ is the orthogonal projector onto $\mathrm{col}(\Phi)^\perp$, and r is the orthogonal projection of t onto that complement.

Proof. Using $\hat{w} = (\Phi^\top\Phi)^{-1}\Phi^\top t$ gives $\hat{t} = \Phi\hat{w} = \Phi(\Phi^\top\Phi)^{-1}\Phi^\top t = Pt$, so $r = t - \hat{t} = (I - P)t$. The identities $P^2 = P$ and $P^\top = P$ follow from straightforward algebra:

$$P^2 = \Phi(\Phi^\top\Phi)^{-1}\underbrace{\Phi^\top\Phi}_{=}(\Phi^\top\Phi)^{-1}\Phi^\top = P, \qquad P^\top = \big(\Phi(\Phi^\top\Phi)^{-1}\Phi^\top\big)^\top = \Phi(\Phi^\top\Phi)^{-1}\Phi^\top = P.$$

Thus P is an orthogonal projector onto $\mathrm{col}(\Phi)$ and $(I - P)$ projects orthogonally onto its complement, so r lies in $\mathrm{col}(\Phi)^\perp$. □

Bayesian Linear Regression: Prior on w and Predictive Distribution

Bayesian Formulation

In Bayesian linear regression we treat the parameter vector w as a random variable and place a prior distribution on it. The generative model is:

$$t = \Phi w + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \beta^{-1}I_N),$$

where β is the noise precision.

Prior Distribution on w

We choose a zero-mean isotropic Gaussian prior:

$$p(w) = \mathcal{N}(w \mid 0, \alpha^{-1}I_M),$$

where α is the prior precision. This encodes the belief that large weights are unlikely (acts as a regularizer).

Likelihood

Conditioned on w, the likelihood of the data is:

$$p(t \mid \Phi, w, \beta) = \mathcal{N}(t \mid \Phi w, \beta^{-1}I_N).$$

Posterior Distribution of w

By Bayes' theorem:

$$p(w \mid t, \Phi) \propto p(t \mid \Phi, w, \beta)\,p(w).$$

Because both prior and likelihood are Gaussian, the posterior is also Gaussian:

$$p(w \mid t, \Phi) = \mathcal{N}(w \mid m_N, S_N),$$

with posterior precision and covariance given by:

$$S_N^{-1} = \alpha I_M + \beta\Phi^\top\Phi, \qquad S_N = (\alpha I_M + \beta\Phi^\top\Phi)^{-1},$$

and the posterior mean:

$$m_N = \beta S_N \Phi^\top t.$$

Interpretation

- m_N is the Bayes estimate of w (posterior mean).
- S_N quantifies uncertainty in the weight estimates.
- As $\alpha \to 0$ (weak prior),

$$m_N \to (\Phi^\top\Phi)^{-1}\Phi^\top t,$$

recovering the ordinary least squares solution.

Predictive Distribution

For a new input x_* with feature vector $\phi_* = \phi(x_*)$, the predictive distribution integrates over the posterior uncertainty in w:

$$p(t_* \mid x_*, t, \Phi) = \int p(t_* \mid x_*, w, \beta)\, p(w \mid t, \Phi)\, dw.$$

The integrand is a product of two Gaussians, so the predictive distribution is Gaussian:

$$p(t_* \mid x_*, t, \Phi) = \mathcal{N}\big(t_* \mid m_N^\top \phi_*,\ \beta^{-1} + \phi_*^\top S_N \phi_*\big).$$

Predictive Mean and Variance

Predictive Mean:

$$\mathbb{E}[t_* \mid x_*, t, \Phi] = m_N^\top \phi_*.$$

Predictive Variance:

$$\mathrm{Var}(t_* \mid x_*, t, \Phi) = \underbrace{\beta^{-1}}_{\text{noise variance}} + \underbrace{\phi_*^\top S_N \phi_*}_{\text{model uncertainty}}.$$

Thus the predictive variance decomposes into:

- aleatoric noise (irreducible), and
- epistemic uncertainty (reduced with more data).

Likelihood Derivation (Gaussian Noise) and MLEs

1. Single-observation likelihood

Assume the data generation model for a single observation:

$$t_n = w^\top \phi(x_n) + \varepsilon_n, \qquad \varepsilon_n \sim \mathcal{N}(0, \beta^{-1}).$$

Then the conditional density (likelihood) for t_n given w is

$$p(t_n \mid x_n, w, \beta) = \mathcal{N}\big(t_n \mid w^\top \phi(x_n),\ \beta^{-1}\big) = \sqrt{\frac{\beta}{2\pi}}\, \exp\Big(-\tfrac{\beta}{2}\big(t_n - w^\top \phi(x_n)\big)^2\Big).$$

2. Joint likelihood for the dataset

Assuming i.i.d. noise, the joint likelihood for all N observations is the product

$$p(t \mid \Phi, w, \beta) = \prod_{n=1}^{N} p(t_n \mid x_n, w, \beta) = \left(\frac{\beta}{2\pi}\right)^{N/2} \exp\Big(-\tfrac{\beta}{2}\sum_{n=1}^{N}(t_n - w^\top \phi(x_n))^2\Big).$$

Using matrix notation with $\Phi \in \mathbb{R}^{N\times M}$ and $t \in \mathbb{R}^N$:

$$p(t \mid \Phi, w, \beta) = \left(\frac{\beta}{2\pi}\right)^{N/2} \exp\Big(-\tfrac{\beta}{2}\|t - \Phi w\|^2\Big).$$

3. Log-likelihood

The log-likelihood (more convenient for optimization) is

$$\ell(w, \beta) := \log p(t \mid \Phi, w, \beta) = \frac{N}{2}\log\beta - \frac{N}{2}\log(2\pi) - \frac{\beta}{2}\|t - \Phi w\|^2.$$

Dropping constants independent of the parameters when optimizing:

$$\ell(w, \beta) = \frac{N}{2}\log\beta - \frac{\beta}{2}\|t - \Phi w\|^2 + \text{const}.$$

4. MLE for w (given β)

Take gradient of the log-likelihood w.r.t. w:

$$\nabla_w \ell(w, \beta) = -\frac{\beta}{2}\cdot 2\,(-\Phi^\top)(t - \Phi w) = \beta\Phi^\top(t - \Phi w).$$

Set to zero for critical point:

$$\Phi^\top(t - \Phi w) = 0 \quad \Rightarrow \quad \Phi^\top\Phi\, w = \Phi^\top t.$$

If $\Phi^\top\Phi$ is invertible, the MLE of w is

$$\boxed{\hat{w}_{\text{MLE}} = (\Phi^\top\Phi)^{-1}\Phi^\top t}$$

which is the ordinary least squares solution. Thus MLE = least squares under Gaussian noise.

5. MLE for noise precision β (given w)

Differentiate ℓ w.r.t. β:

$$\frac{\partial \ell}{\partial \beta} = \frac{N}{2\beta} - \frac{1}{2}\|t - \Phi w\|^2.$$

Set equal to zero:

$$\frac{N}{2\beta} = \frac{1}{2}\|t - \Phi w\|^2 \quad \Rightarrow \quad \hat{\beta}_{\text{MLE}} = \frac{N}{\|t - \Phi w\|^2}.$$

If we substitute $w = \hat{w}_{\text{MLE}}$ we get the MLE for β:

$$\boxed{\hat{\beta}_{\text{MLE}} = \frac{N}{\|t - \Phi\hat{w}_{\text{MLE}}\|^2}}.$$

Equivalently, the MLE for noise variance $\sigma^2 = \beta^{-1}$ is

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{N}\|t - \Phi\hat{w}_{\text{MLE}}\|^2.$$

(For an unbiased estimator of σ^2 divide by $N - M$ instead of N.)

6. Negative log-likelihood and connection to MAP

The negative log-likelihood (up to additive constant) is

$$-\ell(w, \beta) \propto \frac{\beta}{2}\|t - \Phi w\|^2 - \frac{N}{2}\log\beta.$$

When combining with a Gaussian prior $p(w) \propto \exp\big(-\frac{\alpha}{2}\|w\|^2\big)$, the negative log-posterior (up to constants) becomes

$$-\log p(w \mid t) \propto \frac{\beta}{2}\|t - \Phi w\|^2 + \frac{\alpha}{2}\|w\|^2,$$

whose minimizer yields the MAP estimator. Dividing through by β and setting $\lambda = \alpha/\beta$ gives the familiar ridge form:

$$\hat{w}_{\text{MAP}} = (\Phi^\top\Phi + \lambda I)^{-1}\Phi^\top t.$$

Derivation of the Posterior with a Gaussian Prior (Completing the Square)

Assume the Gaussian likelihood and Gaussian prior:

$$p(t \mid w) \propto \exp\Big(-\tfrac{\beta}{2}\|t - \Phi w\|^2 \Big), \qquad p(w) \propto \exp\Big(-\tfrac{\alpha}{2}\|w\|^2 \Big).$$

Posterior (unnormalized) by Bayes' rule:

$$p(w \mid t) \propto p(t \mid w)\, p(w) \propto \exp\Big(-\tfrac{\beta}{2}\|t - \Phi w\|^2 - \tfrac{\alpha}{2}\|w\|^2 \Big).$$

Expand the exponents (quadratic form in w).

$$\begin{aligned} &\tfrac{\beta}{2}\|t - \Phi w\|^2 + \tfrac{\alpha}{2}\|w\|^2 \\ &= \tfrac{\beta}{2}\big(t^\top t - 2t^\top \Phi w + w^\top \Phi^\top \Phi w\big) + \tfrac{\alpha}{2} w^\top w \\ &= \tfrac{1}{2}\, w^\top (\beta \Phi^\top \Phi + \alpha I)\, w \;-\; \beta t^\top \Phi w + \tfrac{\beta}{2} t^\top t. \end{aligned}$$

Group terms in w and complete the square. Write the quadratic form as

$$\tfrac{1}{2}\, w^\top A\, w - b^\top w + \text{const}, \quad \text{where } A = \beta \Phi^\top \Phi + \alpha I, \quad b = \beta \Phi^\top t.$$

Complete the square:

$$\tfrac{1}{2} w^\top A w - b^\top w = \tfrac{1}{2}(w - A^{-1}b)^\top A (w - A^{-1}b) - \tfrac{1}{2} b^\top A^{-1} b.$$

Thus the unnormalized posterior becomes

$$p(w \mid t) \propto \exp\Big(-\tfrac{1}{2}(w - A^{-1}b)^\top A (w - A^{-1}b) \Big) \cdot \exp\Big(\tfrac{1}{2} b^\top A^{-1} b - \tfrac{\beta}{2} t^\top t \Big).$$

The second exponential is independent of w and becomes part of the normalizing constant.

Identify posterior covariance and mean. Hence the posterior is Gaussian with precision A and covariance $S_N = A^{-1}$:

$$S_N = (\beta \Phi^\top \Phi + \alpha I)^{-1},$$

and posterior mean

$$m_N = A^{-1} b = (\beta \Phi^\top \Phi + \alpha I)^{-1} (\beta \Phi^\top t).$$

Simplify using $\lambda = \alpha/\beta$. Dividing numerator and denominator by β gives the more familiar form:

$$S_N = \beta^{-1} (\Phi^\top \Phi + \lambda I)^{-1}, \qquad m_N = (\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top t,$$

where $\lambda = \alpha/\beta$. Note that m_N equals the ridge/MAP estimator and S_N quantifies posterior uncertainty.