S.NO. 2 CHEATSHEET

# Derive the loss from the likelihood

The loss is not an arbitrary penalty: it is the negative log-score of a predictive distribution.

**Distribution assumption → per-example NLL → familiar training loss.**

## 1 · The universal derivation

Conditional independence gives

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_i p_{\theta(y_i|\boldsymbol{x}_i)}.$$

Taking logs and changing maximization to minimization gives

$$\hat{\theta}_{\text{MLE}} = \arg\min_{\theta} -\sum_i \log p_{\theta(y_i|\boldsymbol{x}_i)}.$$

Define one-example loss

$$\ell_{i(\theta)} = -\log p_{\theta(y_i|\boldsymbol{x}_i)}.$$

The dataset objective is the sum or mean of these contributions.

## 2 · Gaussian likelihood gives squared error

Assume

$$Y_i|\boldsymbol{x}_i \sim \mathcal{N}(\mu_i, \sigma^2), \quad \mu_i = f_{\theta(\boldsymbol{x}_i)}.$$

Then

$$-\log p(y_i|\boldsymbol{x}_i) = \tfrac{1}{2\sigma^2}(y_i - \mu_i)^2 + \log\sigma + C.$$

With fixed $\sigma$,

$$\hat{\theta}_{\text{MLE}} = \arg\min_{\theta} \sum_i \left(y_i - f_{\theta(\boldsymbol{x}_i)}\right)^2.$$

MSE is Gaussian NLL up to positive scale and additive constants. Learning $\sigma$ requires keeping the $\log\sigma$ term.

## 3 · Residual model chooses the penalty

Let $r = y - \hat{y}$.

| Noise | NLL in $r$ | Effect |
|---|---|---|
| Gaussian | $r^2$ | large errors dominate |
| Laplace | $\lvert r\rvert$ | constant influence |
| Student-$t$ | $\log\left(1 + \frac{r^2}{\nu s^2}\right)$ | heavy-tail robustness |
| Uniform | $0$ inside; $\infty$ outside | hard support |

Constants and positive scale factors are omitted in the compact table.

## 4 · Bernoulli likelihood gives BCE

For $y \in \{0, 1\}$ and predicted $p = P(Y = 1|\boldsymbol{x})$,

$$p(y|\boldsymbol{x}) = p^{y(1-p)^{1-y}}.$$

Therefore

$$\ell_{\text{BCE}} = -[y\log p + (1-y)\log(1-p)].$$

If $y = 1$, $p = 0.8$ gives $\ell \approx 0.22$ while $p = 0.2$ gives $\ell \approx 1.61$.

Use a logit $z \in \mathbb{R}$ and a fused BCE-with-logits operation for numerical stability.

## 5 · Categorical likelihood gives cross-entropy

For mutually exclusive classes, logits $\boldsymbol{z}$ define

$$p_k = \frac{\exp z_k}{\sum_j \exp z_j}.$$

If the observed class is $c$,

$$\ell_{\text{CE}} = -\log p_c = -z_c + \log\sum_j \exp z_j.$$

Cross-entropy depends on the probability assigned to the true class. Confident mistakes are expensive because $-\log p_c \to \infty$ as $p_c \to 0$.

## 6 · Match observation, output and loss

| Target | Model output | NLL |
|---|---|---|
| real | $\mu$; optionally $\sigma > 0$ | Gaussian / robust |
| binary | one unconstrained logit | Bernoulli BCE |
| one of $K$ | $K$ unconstrained logits | categorical CE |
| count | positive rate $\lambda$ | Poisson NLL |

**Before optimizing a loss, say aloud which random variable and distribution make it the right score.**

S.NO. 2 CHEATSHEET

# From MLE to MAP and regularization

Likelihood asks for data fit; a prior adds an explicit preference among plausible parameter settings.

## 1 · Bayes combines data and prior

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}.$$

| | |
|---|---|
| likelihood | how the fixed data score each $\theta$ |
| prior | pre-data preference over $\theta$ |
| posterior | updated distribution after observing data |
| evidence | normalizer; independent of $\theta$ |

## 2 · MAP is loss plus regularizer

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{D}).$$

Dropping the evidence and taking negative logs,

$$\hat{\theta}_{\text{MAP}} = \arg\min_{\theta} \Big[ \underbrace{-\log p(\mathcal{D}|\theta)}_{\text{NLL}} + \underbrace{-\log p(\theta)}_{\text{prior penalty}} \Big].$$

**MLE** uses only the likelihood.
**MAP** returns one parameter vector at the posterior mode.

## 3 · Gaussian prior gives $L_2$

If $\theta \sim \mathcal{N}(\mathbf{0}, \tau^2 \boldsymbol{I})$,

$$-\log p(\theta) = \tfrac{1}{2\tau^2}\|\theta\|_2^2 + C.$$

Hence

$$\mathcal{L}_{\text{MAP}} = \text{NLL} + \lambda\|\theta\|_2^2, \quad \lambda = \tfrac{1}{2\tau^2}.$$

Smaller $\tau$ means a narrower prior, larger $\lambda$, and stronger shrinkage toward zero.

## 4 · Laplace prior gives $L_1$

For independent zero-centred Laplace parameters with scale $a$,

$$-\log p(\theta) = \tfrac{1}{a}\|\theta\|_1 + C.$$

The corner at zero can threshold coefficients exactly to zero. $L_2$ smoothly shrinks; $L_1$ can create sparsity.

## 5 · One-dimensional MAP shrinkage

Approximate the likelihood by

$$p(\mathcal{D}|\theta) \propto \exp\left(-\tfrac{(\theta - z)^2}{2s^2}\right)$$

with $z = \hat{\theta}_{\text{MLE}}$, and use $\theta \sim \mathcal{N}(0, \tau^2)$. Minimizing the two quadratic terms gives

$$\hat{\theta}_{\text{MAP}} = \underbrace{\tfrac{\tau^2}{\tau^2 + s^2}}_{\alpha \in (0,1)} z.$$

If $z = 0.6$ and $s = \tau$, then $\hat{\theta}_{\text{MAP}} = 0.3$.

## 6 · Scaling: keep the statistical meaning

The exact MAP objective uses **summed** NLL plus the full prior penalty:

$$\textstyle\sum_i \ell_{i(\theta)} + \lambda_{\text{sum}} R(\theta).$$

If code uses mean NLL, divide the whole objective by $n$:

$$\tfrac{1}{n}\textstyle\sum_i \ell_{i(\theta)} + \lambda_{\text{mean}} R(\theta), \quad \lambda_{\text{mean}} = \tfrac{\lambda_{\text{sum}}}{n}.$$

Changing batch size should not silently change the intended regularization strength. Document whether a library returns a sum or mean.

## 7 · What stronger priors do

| | |
|---|---|
| **More data** | likelihood sharpens; MAP often approaches MLE |
| **Narrower prior** | posterior mode moves farther toward preferred values |
| **Correlated prior** | preferred directions become elliptical, not axis-aligned |
| **Prior centre** $m$ | shrinkage is toward $m$, not automatically toward zero |

## 8 · Final audit

- ✓ Each loss is tied to an explicit likelihood.
- ✓ Constants are dropped only when they do not affect the optimized parameters.
- ✓ Logits and constrained distribution parameters are not confused.
- ✓ Prior family and scale are stated.
- ✓ Sum-versus-mean scaling is explicit.

**Likelihood gives the data loss. Prior gives the regularizer. Their posterior mode gives MAP.**