S.NO. 1 CHEATSHEET

From outcomes to predictive distributions

A model should say which values are possible and how plausible each one is.

A neural network maps an input to the parameters of a distribution for the target.

1 · Start with the random variable

The **support** $\mathcal{S}_Y$ is the set of values the target can take. The support should match the task before you choose a loss.

BINARY	$Y \in \{0, 1\}$
CLASS	$Y \in \{1, \ldots, K\}$
COUNT	$Y \in \{0, 1, 2, \ldots\}$
REAL	$Y \in \mathbb{R}$

Impossible targets should receive zero probability or density. A support mismatch is a modelling error, not an optimizer problem.

2 · Mass for discrete, density for continuous

PMF	$p_{Y(y)} = P(Y = y)$ and $\sum_{y \in \mathcal{S}_Y} p_{Y(y)} = 1.$
PDF	$p_{Y(y)} \geq 0$ and $\int_{\mathcal{S}_Y} p_{Y(y)} \, \mathrm{d}y = 1.$

For continuous Y,

$$P(a \leq Y \leq b) = \int_a^b p_{Y(y)} \, \mathrm{d}y.$$

A density value can exceed 1; it is the **area** that is a probability.

3 · Four distribution families

Family	Parameters	Support
Bernoulli	$p \in [0, 1]$	$\{0, 1\}$
Categorical	$p_1, \ldots, p_K$	$\{1, \ldots, K\}$
Poisson	$\lambda > 0$	nonnegative integers
Gaussian	$\mu \in \mathbb{R}, \sigma > 0$	$\mathbb{R}$

4 · Read conditional model notation

$$p_{\boldsymbol{\theta}}(y \mid \boldsymbol{x})$$

$\boldsymbol{x}$	one observed input
y	one possible or observed target
$\boldsymbol{\theta}$	all learned weights and biases
$p_{\theta(y\|\boldsymbol{x})}$	probability or density assigned to y, given $\boldsymbol{x}$

During supervised training the observed inputs are treated as fixed; the model describes the conditional target distribution.

5 · The network predicts distribution parameters

REG	$\boldsymbol{x} \to \left(\mu_{\theta(\boldsymbol{x})}, \sigma_{\theta(\boldsymbol{x})}\right)$
BINARY	$\boldsymbol{x} \to p_{\theta(\boldsymbol{x})}$
K-CLASS	$\boldsymbol{x} \to \left(p_{\theta,1}(\boldsymbol{x}), \ldots, p_{\theta,K}(\boldsymbol{x})\right)$

A point prediction is only a summary: often the conditional mean or mode. Training scores the **full distribution** at the observed target.

6 · Prediction and likelihood ask different questions

Prediction: fix $\boldsymbol{\theta}$ and $\boldsymbol{x}$; vary the possible outcome y.

Likelihood: fix the observed data; vary candidate parameters $\boldsymbol{\theta}$.

The same expression $p_{\theta(y|\boldsymbol{x})}$ is read differently because a different quantity is allowed to vary.

Choose the target variable first, then its support, then a distribution family, then the neural outputs that parameterize it.

S.NO. 1 CHEATSHEET

Build a dataset likelihood

Likelihood ranks parameter settings by how well they explain the outcomes that were actually observed.

1 · Score one fixed observation

For one pair $(\boldsymbol{x}^\star, y^\star)$, the likelihood contribution is

$$L_{i(\boldsymbol{\theta})} = p_{\boldsymbol{\theta}}(y^\star|\boldsymbol{x}^\star).$$

A larger value means that candidate $\boldsymbol{\theta}$ explains this fixed observation better.

For a continuous target this is a density. Comparing candidates is valid even though the density is not itself an event probability.

2 · Independence gives a product

For conditionally independent examples,

$$L(\boldsymbol{\theta}) = p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i).$$

Independent means multiply the per-example factors.
Identically distributed means reuse the same conditional model form.

These are separate assumptions; either can fail without the other.

3 · Tiny Bernoulli comparison

Compare $\theta = 0.5$ and $\theta = 0.8$ where $P(H|\theta) = \theta$.

Data	$L(0.5)$	$L(0.8)$
H, T	$0.5(0.5) = 0.25$	$0.8(0.2) = 0.16$
H, H	$0.5^2 = 0.25$	$0.8^2 = 0.64$

The candidate ranking changes when the fixed observed data change.

4 · MLE names the winner

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}).$$

Maximum likelihood estimation chooses the parameter vector that assigns the largest likelihood to the observed dataset.

5 · Logs preserve the ranking

Because $\log$ is strictly increasing on positive values,

$$\arg\max_{\theta} L(\theta) = \arg\max_{\theta} \log L(\theta).$$

For independent examples, the product becomes a sum:

$$\log L(\boldsymbol{\theta}) = \sum_i \log p_{\theta(y_i|\boldsymbol{x}_i)}.$$

This avoids underflow and makes derivatives easier to combine.

6 · Negative log-likelihood is a loss

Optimizers conventionally minimize, so negate the log-likelihood:

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

$$\hat{R}(\boldsymbol{\theta}) = \tfrac{1}{n}\sum_i \ell_{i(\boldsymbol{\theta})} = -\tfrac{1}{n}\log L(\boldsymbol{\theta}).$$

The factor $\frac{1}{n}$ changes gradient scale, not the minimizer.

7 · Support becomes a hard constraint

If the model assigns $p_{\theta(y_i|\boldsymbol{x}_i)} = 0$ to an observed target, then

$$\ell_i = -\log 0 = +\infty.$$

A zero-likelihood observation is an infinite penalty. Check support before blaming numerical instability.

8 · Final audit

- ✓ Target support matches the chosen distribution.
- ✓ Neural outputs obey parameter constraints.
- ✓ Each observation contributes one log-density or log-probability.
- ✓ Independence assumptions are stated, not silently assumed.
- ✓ Likelihood and prediction are not confused.

Probability models possible outcomes. Likelihood scores parameters using the outcome that occurred.