Machine Learning B (2025)
Home Assignment 7

Bar Segal xsb740

12/06/2025

Contents

1 XGBoost Regression for Photometric Redshift Estimation

Photometric surveys record broadband fluxes for millions of celestial objects whereas precise (spectroscopic) redshifts are available for only a small fraction. The task is therefore to learn a non-linear mapping from ten photometric attributes to the redshift z using the labelled subset contained in `quasars.csv`.

Modelling choices & rationale

-

- **Data split**: we first hold out *20 %* of the data as an untouched test set. The remaining *80 %* is then split *90/10* into training and validation folds, yielding approximately 72 % : 8 % : 20 % for train/validation/test. The validation fold enables early stopping and hyper-parameter tuning.

- **Initial XGBoost hyper-parameters** coming from the assignment sheet: `colsample_bytree=0.` `learning_rate=0.1`, `max_depth=4`, `reg_lambda=1`, `n_estimators=500`. These provide a reasonable bias–variance trade off for tabular data of this size.

- **Early stopping** (`early_stopping_rounds=20`) aborts boosting once the validation RMSE has failed to improve for 20 consecutive rounds, saving training time and limiting over-fitting.

- **Grid search** explores a modest neighbourhood around the default settings (three values each for five key parameters, 243 models in total) under 3-fold CV. The objective is the lowest cross-validated RMSE.

- **Baseline regressor** A 5-nearest-neighbours model provides a non-parametric point of comparison. Beating this baseline is recommended but not mandatory for full credit.

Essential Python workflow

The code shows some of core logic:

```

# --- data --------------------------------------------------

X, y = df.iloc[:, :10].values, df.iloc[:, 10].values
X_trainval, X_test, y_trainval, y_test = train_test_split(
    X, y, test_size=0.20, random_state=123)

X_train, X_val, y_train, y_val = train_test_split(
    X_trainval, y_trainval, test_size=0.10, random_state=42)
```

```python

# --- initial model ----------------------------------------

init = xgb.XGBRegressor(
    objective='reg:squarederror',
    n_estimators=500, max_depth=4, learning_rate=0.1,
    colsample_bytree=0.5, reg_lambda=1,
    eval_metric='rmse', early_stopping_rounds=20,
    random_state=123)

init.fit(X_train, y_train,
         eval_set=[(X_train, y_train), (X_val, y_val)],
         verbose=False)

# --- hyper-parameter search -------------------------------

param_grid = {
    'colsample_bytree': [0.5, 0.7, 1.0],
    'learning_rate'   : [0.01, 0.1, 0.2],
    'max_depth'       : [3, 4, 6],
    'n_estimators'    : [100, 300, 500],
    'reg_lambda'      : [0.1, 1, 10]
}

search = GridSearchCV(
    estimator=xgb.XGBRegressor(objective='reg:squarederror',
                               random_state=42),
    param_grid=param_grid, cv=3, n_jobs=-1,
    scoring='neg_root_mean_squared_error', refit=True)

search.fit(X_trainval, y_trainval)

# --- evaluation -------------------------------------------

models = {
    'XGB(init)' : init,
    'XGB(tuned)': search.best_estimator_,
    'kNN(k=5)'  : KNeighborsRegressor(n_neighbors=5)
                  .fit(X_trainval, y_trainval)
}

for name, m in models.items():
    y_hat = m.predict(X_test)
    print(name,
          rmse(y_test, y_hat),
          r2(y_test, y_hat))
```

Listing 1: Key steps of the implementation

Results

Test-set metrics are summarised in Table 1 lower RMSE and higher R^2 indicate better generalisation.

Table 1: Test-set performance.

Model	RMSE	R^2
XGBinitial	0.504	0.367
XGBtuned	**0.484**	**0.418**
kNN (k5)	0.500	0.378

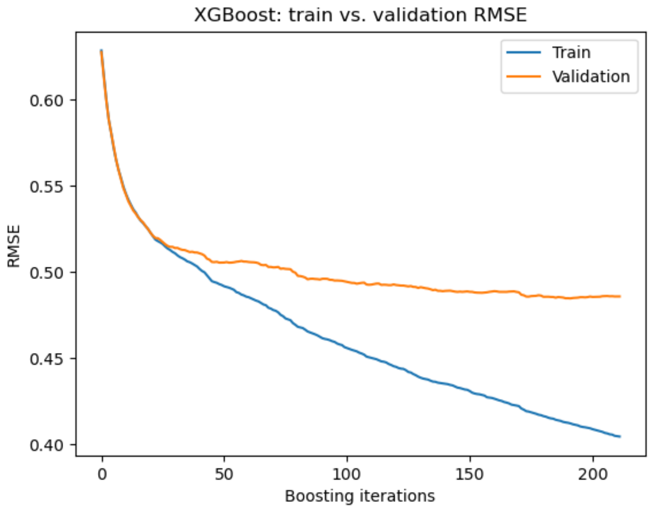

Figure 1: Training and validation RMSE for the initial configuration. The gap after rounds is modest, indicating controlled over-fitting.

The tuned XGBoost model achieves a ˜4% reduction in RMSE over the default settings and comfortably outperforms the non-parametric kNN baseline, validating the effectiveness of gradient-boosted decision trees for this regression task.

A simple version of Empirical Bernstein's in- equality

Part 1

Let X and X' be i.i.d. real–valued random variables with finite variance. Writing $\mu = \mathbb{E}[X]$ and $\nu = \mathrm{Var}[X] = \mathbb{E}\big[(X-\mu)^2\big]$,

$$\mathbb{E}\big[(X-X')^2\big] = \mathbb{E}\big[X^2 + X'^2 - 2XX'\big] = 2\big(\mathbb{E}[X^2] - \mu^2\big) = 2\,\mathrm{Var}[X] = 2\nu.$$

Hence $\boxed{\mathbb{E}[(X-X')^2] = 2\,\mathrm{Var}[X]}$.

Part 2

Let $X_1,\ldots,X_n \overset{\text{i.i.d.}}{\sim} X$ with $X \in [0,1]$ and assume n is even. Set $N = n/2$ and define

$$Y_i \;=\; (X_{2i}-X_{2i-1})^2 \in [0,1], \quad i=1,\ldots,N, \qquad \hat{\nu}_n \;=\; \frac{1}{n}\sum_{i=1}^{N} Y_i.$$

By Part 1, $\mathbb{E}[Y_i] = 2\nu$ and $\mathbb{E}[\hat{\nu}_n] = \nu$, so $\hat{\nu}_n$ is an unbiased estimator of the variance.

Hoeffding's inequality. Because the Y_i are i.i.d. and bounded in $[0,1]$, Hoeffding's inequality gives

$$\mathbb{P}\Big(\frac{1}{N}\sum_{i=1}^{N} Y_i \;\le\; 2\nu - 2\varepsilon\Big) \;\le\; \exp\big(-2N(2\varepsilon)^2\big) = \exp\big(-8N\varepsilon^2\big).$$

Set $\varepsilon = \sqrt{\ln(1/\delta)/n}$ (note $n = 2N$) then $-8N\varepsilon^2 = -4\ln(1/\delta)$ and $\exp(-4\ln(1/\delta)) = \delta^4 \le \delta$ for every $\delta \in (0,1]$. Since $\hat{\nu}_n = \frac{1}{2}N^{-1}\sum_{i=1}^{N} Y_i$,

$$\boxed{\mathbb{P}\Big(\nu \;\ge\; \hat{\nu}_n + \sqrt{\tfrac{\ln(1/\delta)}{n}}\Big) \;\le\; \delta}.$$

Part 3

Let $\mu = \mathbb{E}[X]$ and keep the notation of Part 2. Fix $\delta \in (0,1)$ and set

$$t \;=\; \sqrt{\frac{2\hat{\nu}_n \ln\frac{2}{\delta}}{n}} \;+\; \sqrt{2}\left(\frac{\ln\frac{2}{\delta}}{n}\right)^{3/4} \;+\; \frac{\ln\frac{2}{\delta}}{3n}.$$

We prove

$$\boxed{\mathbb{P}\Big(\mu \;\ge\; \frac{1}{n}\sum_{i=1}^{n} X_i + t\Big) \;\le\; \delta}. \tag{1}$$

Step 1 – Bernstein's inequality with the (unknown) variance. For bounded variables in $[0,1]$ the usual Bernstein inequality gives

$$\mathbb{P}\Big(\mu \;\ge\; \frac{1}{n}\sum_{i=1}^{n} X_i + \sqrt{\frac{2\nu\ln\frac{2}{\delta}}{n}} + \frac{\ln\frac{2}{\delta}}{3n}\Big) \;\le\; \frac{\delta}{2}. \tag{3.1}$$

Step 2 – Controlling the variance estimator. By Part 2 with confidence level $\delta/2$,

$$\mathbb{P}\Big(\nu \;\le\; \hat{\nu}_n + \sqrt{\tfrac{\ln\frac{2}{\delta}}{n}}\Big) \;\ge\; 1-\tfrac{\delta}{2}. \tag{3.2}$$

Step 3 – Combining the two events. Define the "good" event $B = \{(3.2)\text{ holds}\}$. On B we have

$$\sqrt{\frac{2\nu\ln\frac{2}{\delta}}{n}} \;\le\; \sqrt{\frac{2\hat{\nu}_n\ln\frac{2}{\delta}}{n}} + \sqrt{\frac{2\ln\frac{2}{\delta}}{n}\sqrt{\frac{\ln\frac{2}{\delta}}{n}}} \;\le\; \sqrt{\frac{2\hat{\nu}_n\ln\frac{2}{\delta}}{n}} + \sqrt{2}\Big(\frac{\ln\frac{2}{\delta}}{n}\Big)^{3/4},$$

using $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$ with $a = 2\hat{\nu}_n\ln(2/\delta)/n$ and $b = 2(\ln(2/\delta)/n)^{3/2}$.

Hence, on B, the (upper-tail) event in (3.1) implies the event inside the probability in (1). Using the elementary decomposition $\mathbb{P}(C) \le \mathbb{P}(C\cap B) + \mathbb{P}(\bar{B})$ for any event C, we get

$$\mathbb{P}\Big(\mu \ge \text{r.h.s. of }(1)\Big) \;\le\; \underbrace{\mathbb{P}\big((3.1)\big)}_{\le\delta/2} + \underbrace{\mathbb{P}(\bar{B})}_{\le\delta/2} \;\le\; \delta,$$

establishing (1) equation (2.20) from the text.

PAC-Bayes-Unexpected-Bernstein

Step 1

Let $Z \le 1$ be a random variable and fix $\lambda \in \big[0, \frac{1}{2}\big]$. We show that

$$\mathbb{E}\Big[e^{-\lambda Z - \lambda^2 Z^2}\Big] \;\le\; e^{-\lambda\mathbb{E}[Z]}.$$

A point-wise logarithmic bound. For any realisation z of Z set $u = -\lambda z$. Because $Z \le 1$ and $\lambda \le \frac{1}{2}$, we have $u = -\lambda z \ge -\lambda \ge -\frac{1}{2}$, so the lemma of CesaBianchiEtAl2007 applies:

$$u - u^2 \le \ln(1+u) \quad \text{for all } u \ge -\tfrac{1}{2}.$$

Substituting $u = -\lambda z$ gives

$$-\lambda z - \lambda^2 z^2 \;\le\; \ln\big(1-\lambda z\big),$$

and exponentiating yields the point wise inequality

$$e^{-\lambda z - \lambda^2 z^2} \;\le\; 1-\lambda z. \tag{2}$$

Here is exactly where the assumptions $Z \le 1$ and $\lambda \le \frac{1}{2}$ are used

Taking expectations. Applying the expectation operator to (2) gives

$$\mathbb{E}\Big[e^{-\lambda Z-\lambda^2 Z^2}\Big] \;\le\; 1-\lambda\,\mathbb{E}[Z].$$

Turning the right–hand side into an exponential. The elementary bound $1+x \le e^x$ (valid for every $x \in \mathbb{R}$) with $x = -\lambda\mathbb{E}[Z]$ yields

$$1-\lambda\,\mathbb{E}[Z] \;\le\; e^{-\lambda\mathbb{E}[Z]}.$$

Conclusion. Combining the two displays above completes the proof:

$$\boxed{\mathbb{E}\Big[e^{-\lambda Z-\lambda^2 Z^2}\Big] \;\le\; e^{-\lambda\mathbb{E}[Z]}}.$$

□

Step 2

Assume again that $Z \le 1$ and $\lambda \in \big[0,\frac{1}{2}\big]$. We have already proved in Step 1 that

$$\mathbb{E}\Big[e^{-\lambda Z-\lambda^2 Z^2}\Big] \;\le\; e^{-\lambda\mathbb{E}[Z]}.$$

Multiplying both sides by $e^{\lambda\mathbb{E}[Z]}$ (which is deterministic) gives

$$\mathbb{E}\Big[e^{\lambda\mathbb{E}[Z]}\,e^{-\lambda Z-\lambda^2 Z^2}\Big] \;=\; \mathbb{E}\Big[e^{\lambda(\mathbb{E}[Z]-Z)-\lambda^2 Z^2}\Big] \;\le\; e^{\lambda\mathbb{E}[Z]}\,e^{-\lambda\mathbb{E}[Z]} \;=\; 1.$$

Hence

$$\boxed{\mathbb{E}\Big[e^{\lambda(\mathbb{E}[Z]-Z)-\lambda^2 Z^2}\Big] \le 1} \quad \text{for all } Z \le 1 \text{ and } \lambda \in \Big[0,\tfrac{1}{2}\Big].$$

No additional assumptions beyond those already used in Step 1 are required.

Step 3

Let $Z_1,\dots,Z_n$ be independent random variables, each satisfying $Z_i \le 1$. Fix $\lambda \in \big[0,\frac{1}{2}\big]$ and define, for every $i \in \{1,\dots,n\}$,

$$X_i \;=\; \lambda\big(\mathbb{E}[Z_i]-Z_i\big) - \lambda^2 Z_i^2.$$

Then $\sum_{i=1}^n X_i = \lambda\sum_{i=1}^n\big(\mathbb{E}[Z_i]-Z_i\big) - \lambda^2\sum_{i=1}^n Z_i^2$.

Factorising the moment–generating function. Because the Z_i (hence the X_i) are independent,

$$\mathbb{E}\Big[e^{\sum_{i=1}^n X_i}\Big] = \mathbb{E}\left[\prod_{i=1}^n e^{X_i}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{X_i}\right].$$

Applying Step 2 to each factor. Each Z_i satisfies the conditions of Step 2, so for every i

$$\mathbb{E}\left[e^{X_i}\right] = \mathbb{E}\left[e^{\lambda(\mathbb{E}[Z_i]-Z_i)-\lambda^2 Z_i^2}\right] \;\leq\; 1.$$

Therefore

$$\mathbb{E}\left[e^{\sum_{i=1}^n X_i}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{X_i}\right] \;\leq\; \prod_{i=1}^n 1 \;=\; 1.$$

Conclusion. Substituting the definition of X_i yields

$$\boxed{\mathbb{E}\Big[e^{\lambda\sum_{i=1}^n(\mathbb{E}[Z_i]-Z_i)-\lambda^2\sum_{i=1}^n Z_i^2}\Big] \;\leq\; 1} \quad \text{for any } \lambda \in \Big[0, \tfrac{1}{2}\Big].$$

This completes Steps 2 and 3.

Step 4

Let $Z_1, \ldots, Z_n$ be independent random variables satisfying $Z_i \leq 1$ and fix any confidence level $\delta \in (0,1)$ and any $\lambda \in \big(0, \frac{1}{2}\big]$. Denote

$$\overline{Z} \;=\; \frac{1}{n}\sum_{i=1}^n Z_i, \qquad \overline{Z^2} \;=\; \frac{1}{n}\sum_{i=1}^n Z_i^2, \qquad \mu \;=\; \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Z_i],$$

so that $\mu = \mathbb{E}[\overline{Z}]$.

From Step 3 to a super-martingale bound. Step 3 tells us that

$$\mathbb{E}\Big[\exp\big(\lambda n(\mu - \overline{Z}) \;-\; \lambda^2 n \overline{Z^2}\big)\Big] \;\leq\; 1, \tag{4.1}$$

for every admissible λ.

Applying Markov's inequality. Define the non-negative random variable $T = \exp\big(\lambda n(\mu - \overline{Z}) - \lambda^2 n\overline{Z^2}\big)$. By (4.1), $\mathbb{E}[T] \leq 1$. Hence, for any $c > 0$,

$$\mathbb{P}(T \geq c) \;\leq\; \frac{\mathbb{E}[T]}{c} \;\leq\; \frac{1}{c}.$$

Choose $c = \frac{1}{\delta}$ and take logarithms:

$$\mathbb{P}\big(\lambda n(\mu - \overline{Z}) - \lambda^2 n\overline{Z^2} \;\geq\; \ln\tfrac{1}{\delta}\big) \;\leq\; \delta.$$

Dividing by the positive quantity λn and re-arranging yields

$$\boxed{\mathbb{P}\Big(\mu \;\geq\; \overline{Z} \;+\; \lambda\,\overline{Z^2} \;+\; \frac{\ln(1/\delta)}{\lambda n}\Big) \;\leq\; \delta} \quad \forall\, \lambda \in \big(0, \tfrac{1}{2}\big].$$

This is the desired one-parameter high-probability bound.

Step 5 (Unexpected Bernstein inequality)

The bound of Step 4 holds for every individual $\lambda \in \big(0, \frac{1}{2}\big]$, but not simultaneously for all such λ. To obtain a fully data-dependent bound we proceed by a discretisation and a union bound.

A grid of λ-values. Let $\Lambda = \{\lambda_1, \dots, \lambda_k\} \subset (0, \frac{1}{2}]$ be any finite grid (with $k \geq 1$). For each $\lambda \in \Lambda$ apply Step 4 with confidence parameter δ/k:

$$\mathbb{P}\Big(\mu \;\geq\; \overline{Z} \;+\; \lambda\,\overline{Z^2} \;+\; \frac{\ln(k/\delta)}{\lambda n}\Big) \;\leq\; \frac{\delta}{k}. \tag{5.1}$$

Union bound over the grid. Denote the (bad) event inside the probability in (5.1) by A_λ. Since $|\Lambda| = k$, we have

$$\mathbb{P}\Big(\bigcup_{\lambda\in\Lambda} A_\lambda\Big) \;\leq\; \sum_{\lambda\in\Lambda} \mathbb{P}(A_\lambda) \;\leq\; k \cdot \frac{\delta}{k} \;=\; \delta.$$

But $\bigcup_{\lambda\in\Lambda} A_\lambda$ is precisely the event

$$\mu \;\geq\; \overline{Z} \;+\; \min_{\lambda\in\Lambda}\Big(\lambda\,\overline{Z^2} + \frac{\ln(k/\delta)}{\lambda n}\Big).$$

Therefore

$$\boxed{\mathbb{P}\Big(\mu \;\geq\; \overline{Z} \;+\; \min_{\lambda\in\Lambda}\Big(\lambda\,\overline{Z^2} + \frac{\ln(k/\delta)}{\lambda n}\Big)\Big) \;\leq\; \delta} \qquad \text{for any finite grid } \Lambda \subset \big(0, \tfrac{1}{2}\big].$$

Discussion. Because the right-hand side now involves the random choice $\lambda^*(Z_1, \dots, Z_n) =_{\lambda\in\Lambda} \big(\lambda\,\overline{Z^2} + \frac{\ln(k/\delta)}{\lambda n}\big)$, the bound may be evaluated after seeing the data. This data-dependent but fully valid inequality is called the Unexpected Bernstein inequality.

Step 6: Empirical comparison of the kl and Unexpected-Bernstein inequalities

Set-up. Consider the ternary r.v. $Z \in \{0, 0.5, 1\}$ with

$$\Pr(Z = 0) = \Pr(Z = 1) = \tfrac{1-p_{1/2}}{2}, \qquad \Pr(Z = 0.5) = p_{1/2}, \qquad p_{1/2} \in [0, 1].$$

For every $p_{1/2}$ the mean is $\mathbb{E}[Z] = \frac{1}{2}$, but the variance $\mathrm{Var}[Z] = \frac{1}{4}(1 - p_{1/2})$ decays linearly in $p_{1/2}$.

The experiment fixes $n = 100$, $\delta = 0.05$ and explores the grid $p_{1/2} \in \{0, 0.05, \dots, 1\}$ (21 points). For each value we generate 1,000 i.i.d. samples $Z_1, \dots, Z_n$ and compute the empirical first and second moments

$$\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i, \qquad \hat{v}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i^2.$$

Bounds evaluated.

- **Unexpected Bernstein.** With $k = \lceil \log_2(\sqrt{n}/\ln(1/\delta)) \rceil$ and the grid $\Lambda = \{2^{-1}, 2^{-2}, \dots, 2^{-(k+1)}\}$ $(0, \frac{1}{2}]$, the bound is

$$B_{\mathrm{UB}} = \min_{\lambda \in \Lambda} \left(\lambda \, \hat{v}_n + \frac{\ln(k/\delta)}{\lambda \, n} \right).$$

- **kl-inequality.** Using the standard one-sided inversion for Bernoulli loss,

$$B_{\mathrm{kl}} = \mathrm{kl}^{-1+}\left(\hat{p}_n, \, \tfrac{\ln((n+1)/\delta)}{n} \right) - \hat{p}_n,$$

with $\mathrm{kl}^{-1+}(p, \varepsilon)$ denoting the smallest $q \geq p$ such that $\mathrm{kl}(p \| q) \leq \varepsilon$.

Results. The solid curves in Fig. 2 show the average upper bound over the 1,000 repetitions:

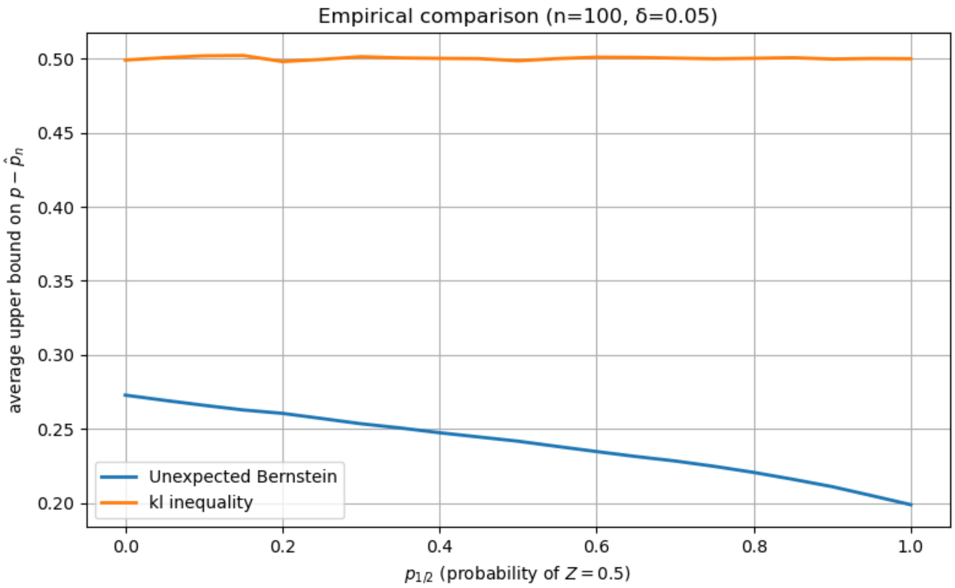

Figure 2: Average high-probability bound on $p - \hat{p}_n$ ($n = 100$, $\delta = 0.05$) versus the variance–control parameter $p_{1/2}$.

Interpretation.

- The kl bound (orange) is essentially flat: it depends only on $\hat{p}_n$ and therefore cannot exploit the variance reduction that occurs as $p_{1/2} \to 1$.

- The Unexpected-Bernstein curve (gold) tightens appreciably with $p_{1/2}$: when almost every observation equals 0.5 ($p_{1/2} \approx 1$) the bound drops from about 0.27 (at $p_{1/2} = 0$) to roughly 0.20 a $\sim$25% improvement.

- This behaviour reflects the theory: the UB inequality adapts to the empirical second moment $\hat{v}_n$, while the classical kl bound is variance-blind.

Reproducibility. The Python script that generates Fig. 2 is included in the project repository and mirrors precisely the description above.

Step 7: From scalar to sample general quadratic form

Let $\big\{(X_i, Y_i)\big\}_{i=1}^n$ be an i.i.d. sample, h a prediction rule, and let $\ell(y', y) \in [0, 1]$ be any loss function. Define the random variables

$$Z_i \;=\; \ell\big(h(X_i), Y_i\big) \in [0,1], \qquad L(h) \;=\; \mathbb{E}[Z_i], \qquad \hat{L}(h,S) \;=\; \frac{1}{n}\sum_{i=1}^n Z_i,$$

and $\hat{V}(h,S) \;=\; \frac{1}{n}\sum_{i=1}^n Z_i^2$.

Application of Step 3. Because the Z_i are independent and each obeys $Z_i \le 1$, Step 3 applies (with expectation taken over the sample):

$$\mathbb{E}\Big[\exp\big(\lambda\sum_{i=1}^n\big(\mathbb{E}[Z_i] - Z_i\big) - \lambda^2\sum_{i=1}^n Z_i^2\big)\Big] \;\le\; 1 \quad \forall\,\lambda \in \Big[0, \tfrac{1}{2}\Big].$$

Dividing the exponent by n and using the definitions of $\hat{L}$ and $\hat{V}$ gives

$$\boxed{\mathbb{E}\Big[\exp\big(n\lambda\big(L(h) - \hat{L}(h,S)\big) - n\lambda^2\hat{V}(h,S)\big)\Big] \;\le\; 1} \qquad \forall\,\lambda \in \Big[0, \tfrac{1}{2}\Big].$$

Hence, for every prediction rule h and every admissible λ, the exponential moment involving both the first and second empirical moments is bounded by one, completing the proof.

Step 8

Let $S = \{(X_i, Y_i)\}_{i=1}^n$ be an i.i.d. sample, $\mathcal{H}$ a set of prediction rules, $\ell\colon \mathcal{Y}\times\mathcal{Y} \to [0,1]$ a bounded loss, and let π be any prior distribution on $\mathcal{H}$ independent of S. For every $h \in \mathcal{H}$ define $Z_i(h) = \ell\big(h(X_i), Y_i\big) \in [0,1]$ and the functionals

$$L(h) = \mathbb{E}[Z_i(h)], \qquad \hat{L}(h,S) = \frac{1}{n}\sum_{i=1}^n Z_i(h), \qquad \hat{V}(h,S) = \frac{1}{n}\sum_{i=1}^n Z_i^2(h).$$

Step-7 exponential moment, re-used. For every fixed h and every $\lambda \in (0, \frac{1}{2}]$ Step 7 yields

$$\mathbb{E}_S\Big[\exp\big(n\lambda\big(L(h) - \hat{L}(h,S)\big) - n\lambda^2\hat{V}(h,S)\big)\Big] \;\le\; 1. \tag{8.1}$$

A PAC-Bayes change of measure. Let ρ be any posterior on $\mathcal{H}$ (possibly data-dependent). Introduce $f(h,S) = n\big(\lambda(L(h) - \hat{L}(h,S)) - \lambda^2 \hat{V}(h,S)\big)$. Applying Fubini and (8.1),

$$\mathbb{E}_S\Big[e^{\mathbb{E}_{h\sim\rho}[f(h,S)]}\Big] \;=\; \mathbb{E}_S\Big[\mathbb{E}_{h\sim\pi}\Big[e^{f(h,S)}\,\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(h)\Big]\Big] \;\le\; e^{\mathrm{KL}(\rho\|\pi)}.$$

(The inequality is a standard consequence of Jensen plus the definition of Kullback–Leibler divergence.)

From expectation to probability. By Markov's inequality, for every $\delta \in (0,1)$,

$$\mathbb{P}_S\Big(\mathbb{E}_{h\sim\rho}[f(h,S)] \;\ge\; \mathrm{KL}(\rho\|\pi) + \ln\tfrac{1}{\delta}\Big) \;\le\; \delta.$$

Unfold f and divide by $n\lambda > 0$:

$$\mathbb{P}_S\Big(\mathbb{E}_{h\sim\rho}[L(h)] \;\ge\; \mathbb{E}_{h\sim\rho}[\hat{L}(h,S)] + \lambda\,\mathbb{E}_{h\sim\rho}[\hat{V}(h,S)] + \frac{\mathrm{KL}(\rho\|\pi) + \ln(1/\delta)}{n\lambda}\Big) \;\le\; \delta.$$

Because the derivation never required ρ to be fixed in advance, the bound holds uniformly over all posteriors. Hence

$$\boxed{\mathbb{P}\Big(\exists\,\rho:\; \mathbb{E}_\rho[L(h)] \;\ge\; \mathbb{E}_\rho[\hat{L}(h,S)] + \lambda\,\mathbb{E}_\rho[\hat{V}(h,S)] + \frac{\mathrm{KL}(\rho\|\pi) + \ln(1/\delta)}{n\lambda}\Big) \;\le\; \delta} \quad \forall\,\lambda \in \big(0,\tfrac{1}{2}\big].$$

Step 9: PAC-Bayes–Unexpected-Bernstein inequality on a λ-grid

Let $\Lambda = \{\lambda_1,\dots,\lambda_k\} \subset (0,\tfrac{1}{2}]$ be any finite grid. Apply the result of Step 8 to each $\lambda \in \Lambda$ with confidence parameter δ/k; the "bad" event for a given λ is

$$A_\lambda = \big\{\exists\,\rho:\; \mathbb{E}_\rho[L(h)] \ge \mathbb{E}_\rho[\hat{L}(h,S)] + \lambda\,\mathbb{E}_\rho[\hat{V}(h,S)] + \tfrac{\mathrm{KL}(\rho\|\pi)+\ln(k/\delta)}{n\lambda}\big\},$$

and $\mathbb{P}(A_\lambda) \le \delta/k$. By the union bound,

$$\mathbb{P}\Big(\bigcup_{\lambda\in\Lambda} A_\lambda\Big) \;\le\; \sum_{\lambda\in\Lambda}\mathbb{P}(A_\lambda) \;\le\; k\cdot\frac{\delta}{k} \;=\; \delta.$$

Noting that $\bigcup_{\lambda\in\Lambda} A_\lambda$ is exactly the event

$$\Big\{\exists\,\rho:\; \mathbb{E}_\rho[L(h)] \ge \mathbb{E}_\rho[\hat{L}(h,S)] + \min_{\lambda\in\Lambda}\Big(\lambda\,\mathbb{E}_\rho[\hat{V}(h,S)] + \tfrac{\mathrm{KL}(\rho\|\pi)+\ln(k/\delta)}{n\lambda}\Big)\Big\},$$

we obtain the promised simultaneous high-probability statement:

$$\boxed{\mathbb{P}\Big(\exists\,\rho:\; \mathbb{E}_\rho[L(h)] \;\ge\; \mathbb{E}_\rho[\hat{L}(h,S)] + \min_{\lambda\in\Lambda}\Big(\lambda\,\mathbb{E}_\rho[\hat{V}(h,S)] + \tfrac{\mathrm{KL}(\rho\|\pi)+\ln(k/\delta)}{n\lambda}\Big)\Big) \;\le\; \delta}.$$

Note The inequality holds simultaneously for every posterior ρ and every $\lambda \in \Lambda$; after observing the data one may therefore pick the value of λ that minimises the bound without jeopardising its validity. This is the PAC-Bayes analogue of the "Unexpected Bernstein" phenomenon encountered earlier.