

The cross-entropy method

Sections 4.1–4.3

Junqi Tang

The proposal controls simulation error

Our target is $I = \mathbb{E}_p[f(X)]$. For $X_i \stackrel{\text{iid}}{\sim} q$,

$$\hat{I}_N = \frac{1}{N} \sum_{i=1}^N H_i, \quad H_i = f(X_i) \frac{p(X_i)}{q(X_i)}.$$

If $q > 0$ wherever $fp \neq 0$ and the second moment is finite,

$$\text{Var}(\hat{I}_N) = \frac{1}{N} \left\{ \int \frac{f(x)^2 p(x)^2}{q(x)} dx - I^2 \right\}.$$

The design problem

Choose a proposal that puts samples where they contribute to the integral, while keeping weighted outputs stable.

The ideal proposal

For this lecture assume $f \geq 0$ and $0 < I = \mathbb{E}_p f < \infty$. Define

$$q^*(x) = \frac{f(x)p(x)}{I}.$$

For $X \sim q^*$, the weighted output is constant almost surely:

$$f(X) \frac{p(X)}{q^*(X)} = I.$$

- This proposal has zero variance.
- Its density contains the unknown answer I .
- We will approximate its shape within a tractable family.

If f changes sign, the variance-optimal density is proportional to $|f|p$; it generally does not give zero variance.

Section 4.1: fitting the ideal distribution

Choose a family $\{q_\theta : \theta \in \Theta\}$ that is easy to sample.

ideal density q^* $\longrightarrow$ best approximation within the family q_θ .

- A richer family can approximate more shapes.
- A simple family may be easier to fit and simulate.
- We need a criterion for comparing q^* and q_θ .

Cross-entropy approach

Choose θ to minimise $D_{\text{KL}}(q^* \| q_\theta)$.

A running two-state example

Let $X \in \{0, 1\}$ with $p(0) = p(1) = 1/2$, and set $f(x) = 1 + 2x$.

x	$p(x)$	$f(x)$	$f(x)p(x)$
0	1/2	1	1/2
1	1/2	3	3/2

Thus $I = 2$ and $q^*(0) = 1/4$, $q^*(1) = 3/4$.

$$q_r = \text{Bernoulli}(r), \quad 0 < r < 1.$$

The ideal proposal samples state 1 more often because its contribution is larger.

Section 4.2: KL divergence

For densities p_1, p_2 , using natural logarithms,

$$D_{\text{KL}}(p_1 \| p_2) = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} dx.$$

For probability masses on a discrete state space S ,

$$D_{\text{KL}}(p_1 \| p_2) = \sum_{x \in S} p_1(x) \log \frac{p_1(x)}{p_2(x)}.$$

Equivalently, it is the expected log density ratio under p_1 :

$$D_{\text{KL}}(p_1 \| p_2) = \mathbb{E}_{p_1} \left[\log \frac{p_1(X)}{p_2(X)} \right].$$

Support and infinite divergence

The conventions are

$$0 \log(0/b) = 0, \quad a \log(a/0) = +\infty \quad (a > 0).$$

If p_1 assigns positive probability to a region where $p_2 = 0$, then

$$D_{\text{KL}}(p_1 \| p_2) = +\infty.$$

Discrete example

If $p_1 = (1/2, 1/2)$ and $p_2 = (1, 0)$, the second state makes the divergence infinite.

Covering the support is necessary for finite KL; it is not sufficient in general. Integrability of the log ratio also matters.

Interpreting the log ratio

At a sampled point x ,

$$\log \frac{p_1(x)}{p_2(x)} = \log p_1(x) - \log p_2(x).$$

- If $p_2(x)$ is much smaller than $p_1(x)$, this contribution is positive and large.
- If $p_2(x) > p_1(x)$, this contribution is negative.
- KL averages these contributions using p_1 .

A property of the average

Individual log ratios may be negative, but their p_1 -weighted average is never negative.

Jensen's inequality

For a convex function ϕ ,

$$\phi(\mathbb{E}[Y]) \leq \mathbb{E}[\phi(Y)]$$

whenever the expressions are well-defined. For positive Y , take

$$\phi(y) = -\log y, \quad \phi''(y) = \frac{1}{y^2} > 0.$$

Then

$$\boxed{\mathbb{E}[-\log Y] \geq -\log \mathbb{E}[Y].}$$

The function $\log y$ is concave; $-\log y$ is strictly convex.

Why KL is nonnegative

Let $S = \{x : p_1(x) > 0\}$ and suppose $p_2 > 0$ p_1 -almost surely. For $X \sim p_1$, set $Y = p_2(X)/p_1(X)$.

$$\begin{aligned} D_{\text{KL}}(p_1 \| p_2) &= \mathbb{E}_{p_1}[-\log Y] \\ &\geq -\log \mathbb{E}_{p_1}[Y] \\ &= -\log \int_S p_2(x) dx \geq 0. \end{aligned}$$

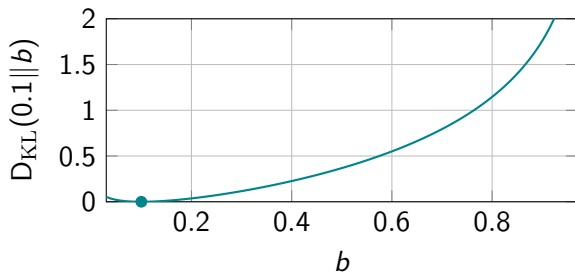
Equality requires Y to be constant and $\int_S p_2 = 1$. Normalisation then gives

$$D_{\text{KL}}(p_1 \| p_2) = 0 \iff p_1 = p_2 \text{ almost everywhere.}$$

KL between Bernoulli distributions

For $p_1 = \text{Bernoulli}(a)$ and $p_2 = \text{Bernoulli}(b)$,

$$D_{\text{KL}}(a\|b) = a \log \frac{a}{b} + (1 - a) \log \frac{1 - a}{1 - b}.$$



The minimum is zero at $b = a = 0.1$.

As the proposal assigns almost no mass to a state that matters under p_1 , KL grows.

Curve: analytic formula evaluated numerically.

KL is not a distance metric

For Bernoulli distributions:

$$D_{\text{KL}}(0.1\|0.5) \approx 0.3681, \quad D_{\text{KL}}(0.5\|0.1) \approx 0.5108.$$

So KL is **not symmetric**. The triangle inequality can also fail:

$$\underbrace{D_{\text{KL}}(0.1\|0.9)}_{1.7578} > \underbrace{D_{\text{KL}}(0.1\|0.5)}_{0.3681} + \underbrace{D_{\text{KL}}(0.5\|0.9)}_{0.5108}.$$

Terminology

KL is a divergence: nonnegative, and zero only for the same distribution. It is not a metric.

KL and cross-entropy

For discrete distributions, define entropy and cross-entropy by

$$H(p_1) = - \sum_x p_1(x) \log p_1(x), \quad H(p_1, p_2) = - \sum_x p_1(x) \log p_2(x).$$

When the quantities are finite,

$$\boxed{D_{\text{KL}}(p_1 \| p_2) = H(p_1, p_2) - H(p_1).}$$

- KL and cross-entropy are different quantities.
- With p_1 fixed, minimising one over p_2 minimises the other.
- Continuous analogues use integrals; differential entropy may be negative.

The direction of KL matters

The CE objective is

$$\min_{\theta \in \Theta} D_{\text{KL}}(q^* \| q_{\theta}).$$

- The expectation is under the *ideal* density q^* .
- It penalises assigning too little proposal density where q^* puts mass.
- Reversing the arguments gives a different optimisation problem.

Support requirement

Every candidate used must cover $\{x : f(x)p(x) > 0\}$, up to null sets.

Section 4.3: the unknown integral cancels

Under suitable integrability, write

$$D_{\text{KL}}(q^* \| q_\theta) = \underbrace{\int q^*(x) \log q^*(x) dx}_{\text{independent of } \theta} - \int q^*(x) \log q_\theta(x) dx.$$

Substitute $q^* = fp/l$:

$$\arg \min_{\theta} D_{\text{KL}}(q^* \| q_\theta) = \arg \max_{\theta} \frac{1}{l} \int f(x)p(x) \log q_\theta(x) dx.$$

Since $l > 0$ is constant in θ ,

$$\theta_{\text{CE}} \in \arg \max_{\theta \in \Theta} F(\theta), \quad F(\theta) = \mathbb{E}_p[f(X) \log q_\theta(X)].$$

CE fitting is weighted maximum likelihood

Draw a pilot sample $X_1, \dots, X_M \stackrel{\text{iid}}{\sim} p$ and maximise

$$\hat{F}_M(\theta) = \frac{1}{M} \sum_{i=1}^M f(X_i) \log q_\theta(X_i).$$

If $\sum_i f(X_i) > 0$, the same maximiser is obtained from

$$\sum_{i=1}^M a_i \log q_\theta(X_i), \quad a_i = \frac{f(X_i)}{\sum_j f(X_j)}.$$

- Ordinary maximum likelihood gives every observation equal weight.
- CE gives more weight to observations with larger $f(X_i)$.
- The unknown I is unnecessary, but the pilot objective is noisy.

Fitting the Bernoulli proposal

For $q_r(x) = r^x(1-r)^{1-x}$, the weighted log likelihood is

$$L(r) = \sum_i a_i \{X_i \log r + (1 - X_i) \log(1 - r)\}.$$

With $\sum_i a_i = 1$, an interior maximiser satisfies

$$\hat{r} = \sum_i a_i X_i.$$

For our running example, $f(0) = 1$ and $f(1) = 3$. If the pilot has n_0 zeros and n_1 ones,

$$\hat{r} = \frac{3n_1}{n_0 + 3n_1}.$$

If a state is absent, the unconstrained fit may lie at 0 or 1. Constrain r away from those boundaries to preserve coverage.

The one-pass CE algorithm

Use separate pilot and production sample sizes, M and N .

1. Draw pilot observations $Z_1, \dots, Z_M \stackrel{\text{iid}}{\sim} p$.
2. Fit $\hat{\theta} \in \arg \max_{\theta} \sum_i f(Z_i) \log q_{\theta}(Z_i)$.
3. Freeze $q_{\hat{\theta}}$ and draw *fresh* $X_1, \dots, X_N \stackrel{\text{iid}}{\sim} q_{\hat{\theta}}$.
4. Return ordinary importance sampling:

$$\hat{I}_N = \frac{1}{N} \sum_{i=1}^N f(X_i) \frac{p(X_i)}{q_{\hat{\theta}}(X_i)}.$$

Also report the sample standard deviation of the H_i divided by $\sqrt{N}$, when a finite-variance approximation is appropriate.

Adaptive fitting needs an importance correction

At iteration k , draw $Z_i \stackrel{\text{iid}}{\sim} q_{\theta_k}$ instead of p . Rewrite

$$F(\theta) = \mathbb{E}_{q_{\theta_k}} \left[f(Z) \frac{p(Z)}{q_{\theta_k}(Z)} \log q_{\theta}(Z) \right].$$

With nonnegative fitting weights

$$\alpha_i = f(Z_i) \frac{p(Z_i)}{q_{\theta_k}(Z_i)},$$

use

$$\theta_{k+1} \in \arg \max_{\theta} \sum_i \alpha_i \log q_{\theta}(Z_i).$$

The denominator uses the proposal that generated the observations.

Adaptive Bernoulli update

If the current proposal is $\text{Bernoulli}(r_k)$, our example has

$$\alpha(0) = \frac{1/2}{1 - r_k}, \quad \alpha(1) = \frac{3/2}{r_k}.$$

For current counts n_0, n_1 ,

$$r_{k+1} = \frac{n_1(3/2)/r_k}{n_0(1/2)/(1 - r_k) + n_1(3/2)/r_k}.$$

At the population level, $n_1/M \rightarrow r_k$ and $n_0/M \rightarrow 1 - r_k$, giving $r_{k+1} \rightarrow 3/4$. Every iteration approximates the same population objective. Iteration changes where samples are gathered; it does not guarantee monotone improvement.

The adaptive CE algorithm

Choose θ_0 , a pilot size M , and a maximum number of updates K .

1. For $k = 0, \dots, K - 1$:

- draw $Z_i \stackrel{\text{iid}}{\sim} q_{\theta_k}$;
- compute $\alpha_i = f(Z_i)p(Z_i)/q_{\theta_k}(Z_i)$;
- check that the weights are finite and have positive sum;
- fit θ_{k+1} by maximising $\sum_i \alpha_i \log q_{\theta}(Z_i)$.

2. Stop early if a chosen diagnostic stabilises; retain the latest θ_K if all K updates are used.

3. Freeze the final proposal, draw fresh production samples, and return

$$\hat{l}_N = \frac{1}{N} \sum_i f(X_i) \frac{p(X_i)}{q_{\text{final}}(X_i)}.$$

Parameter stability is a stopping rule, not a certificate of low error.

When fitting fails

- **No contribution:** all $f(Z_i) = 0$ makes every fitting weight zero; the sample gives no direction for fitting.
- **Too little coverage:** important regions may be absent from the pilot.
- **Unstable fit:** a few observations can dominate the objective or drive parameters to a boundary.
- **Limited family:** the chosen proposal may not represent the important shape.

Practical responses

Inspect pilot coverage and weighted outputs; use support-preserving constraints, more pilot samples, or a better proposal family. Rare events may need staged thresholds (Section 4.5).

KL fitting and variance minimisation

For $f \geq 0$ and $q^* = fp/I$, ordinary IS satisfies

$$\frac{N \text{Var}(\hat{I}_N)}{I^2} = \int \frac{q^*(x)^2}{q_\theta(x)} dx - 1.$$

CE instead minimises

$$D_{\text{KL}}(q^* \| q_\theta) = \int q^*(x) \log \frac{q^*(x)}{q_\theta(x)} dx.$$

- Both achieve their ideal value at $q_\theta = q^*$, if available.
- Within a restricted family, their optimisers may differ.
- Small KL alone does not guarantee small, or even finite, IS variance.

Lecture summary

$$q^* = \frac{fp}{I}, \quad \theta_{\text{CE}} \in \arg \max_{\theta} \mathbb{E}_p[f \log q_{\theta}].$$

- CE approximates an ideal importance-sampling distribution.
- KL is nonnegative and asymmetric; it is not a distance metric.
- The CE objective becomes a weighted maximum-likelihood fit.
- Adaptive fitting corrects for each pilot sampling distribution.
- Fresh production samples give a simple valid final IS estimate.