

Cross Entropy for Gaussian importance sampling

Section 4.4

Junqi Tang

Why fit a Gaussian proposal?

In a simulation model, $f(x)$ may quantify failure, loss, or another contribution. We want

$$I = \int f(x)p(x) dx.$$

A Gaussian proposal gives a tractable way to adjust

- **location:** which input values we visit most often;
- **spread:** how widely we explore around that location.

A computational choice

We change the distribution used to generate simulated inputs, then correct using importance weights. The original model p stays fixed.

The CE objective

Assume $f \geq 0$ and $0 < I = \mathbb{E}_p[f(X)] < \infty$. The ideal IS density is

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

The CE fit maximises

$$F(\theta) = \int f(x)p(x) \log q_\theta(x) dx = I \mathbb{E}_{q^*}[\log q_\theta(X)].$$

- I is a positive constant, so it does not affect the maximiser.
- For the Gaussian formulas, assume finite second moments under q^* .
- We fit a Gaussian to q^* , which need not itself be Gaussian.

Independent Gaussian coordinates

Write $v_j = \sigma_j^2 > 0$. Section 4.4 uses

$$q_{\mu, \nu}(x) = \prod_{j=1}^d \frac{1}{\sqrt{2\pi v_j}} \exp\left[-\frac{(x_j - \mu_j)^2}{2v_j}\right].$$

Equivalently,

$$X \sim \mathcal{N}(\mu, \text{diag}(v_1, \dots, v_d)).$$

- μ_j is the proposal mean of coordinate j .
- v_j is its variance, and $\sqrt{v_j}$ is its standard deviation.
- Independence is a restriction on the proposal family.

The Gaussian log density

Taking logarithms turns the product into a sum:

$$\log q_{\mu, \nu}(x) = -\frac{d}{2} \log(2\pi) - \frac{1}{2} \sum_{j=1}^d \left[\log \nu_j + \frac{(x_j - \mu_j)^2}{\nu_j} \right].$$

Therefore

$$F(\mu, \nu) = C - \frac{1}{2} \sum_{j=1}^d \left[l \log \nu_j + \frac{\int (x_j - \mu_j)^2 f(x) p(x) dx}{\nu_j} \right],$$

where $C = -\frac{dl}{2} \log(2\pi)$ is independent of μ, ν . Each summand depends only on μ_j, ν_j , so we can optimise one coordinate at a time.

A one-coordinate optimisation problem

Define, for coordinate j ,

$$B_j(\mu_j) = \int (x_j - \mu_j)^2 f(x) p(x) dx.$$

The relevant part of the objective is

$$F_j(\mu_j, v_j) = -\frac{I}{2} \log v_j - \frac{B_j(\mu_j)}{2v_j}.$$

- The mean controls weighted squared distance B_j .
- The variance balances concentration against spread.
- A smaller variance does not always increase the objective.

Deriving the mean update

For fixed $v_j > 0$,

$$\frac{\partial F}{\partial \mu_j} = \frac{1}{v_j} \int (x_j - \mu_j) f(x) p(x) dx.$$

Setting this derivative to zero gives

$$\int x_j f(x) p(x) dx - \mu_j I = 0,$$

so

$$\mu_j^* = \frac{\int x_j f(x) p(x) dx}{I}.$$

The second derivative is $-I/v_j < 0$, confirming the unique maximum in μ_j for fixed variance.

Deriving the variance update

With μ_j fixed,

$$\frac{\partial F}{\partial v_j} = -\frac{I}{2v_j} + \frac{B_j(\mu_j)}{2v_j^2}.$$

Multiplying the stationarity equation by $2v_j^2$ gives

$$-Iv_j + B_j(\mu_j) = 0.$$

At the fitted mean,

$$v_j^* = \frac{\int (x_j - \mu_j^*)^2 f(x) p(x) dx}{I}.$$

Equivalently, differentiating in σ_j gives $-I/\sigma_j + B_j/\sigma_j^3 = 0$, with $v_j = \sigma_j^2$.

Why these stationary points are maxima

Let $m_j = \mathbb{E}_{q^*}[X_j]$ and $s_j^2 = \text{Var}_{q^*}(X_j)$. Then

$$\frac{B_j(\mu_j)}{I} = s_j^2 + (\mu_j - m_j)^2.$$

Thus the best mean is $\mu_j = m_j$ for every $v_j > 0$. At that mean,

$$\frac{\partial F_j}{\partial v_j} = \frac{I(s_j^2 - v_j)}{2v_j^2}.$$

If $s_j^2 > 0$, the derivative is positive below s_j^2 and negative above it.

$$\boxed{\mu_j^* = m_j, \quad v_j^* = s_j^2.}$$

Gaussian CE matches moments

Since $q^* = fp/I$,

$$\mu_j^* = \mathbb{E}_{q^*}[X_j], \quad v_j^* = \text{Var}_{q^*}(X_j).$$

- The Gaussian matches the ideal distribution's marginal means and variances.
- These generally differ from the moments under the original p .
- Matching moments does not match every aspect of the distribution.
- A diagonal Gaussian cannot represent correlations or multiple modes.

Estimating the moments using a pilot sample

Draw $Z_1, \dots, Z_M \stackrel{\text{iid}}{\sim} p$ and let

$$a_i = \frac{f(Z_i)}{\sum_{\ell=1}^M f(Z_\ell)}, \quad \sum_i a_i = 1.$$

Provided the denominator is positive,

$$\hat{\mu}_j = \sum_{i=1}^M a_i Z_{ij}, \quad \hat{v}_j = \sum_{i=1}^M a_i (Z_{ij} - \hat{\mu}_j)^2.$$

- i indexes samples and j indexes coordinates.
- Evaluate f on the *whole* vector Z_i .
- Each coordinate uses the same weights a_i .

Pilot samples from a different proposal

If $Z_i \stackrel{\text{iid}}{\sim} q_k$, use the change of measure

$$\int h(x)f(x)p(x) dx = \mathbb{E}_{q_k} \left[h(Z)f(Z) \frac{p(Z)}{q_k(Z)} \right].$$

Define

$$\alpha_i = f(Z_i) \frac{p(Z_i)}{q_k(Z_i)}, \quad a_i = \frac{\alpha_i}{\sum_{\ell} \alpha_{\ell}}.$$

The choices $h(x) = 1$, x_j , and $(x_j - \mu_j)^2$ give the required moment integrals. The generating proposal q_k stays fixed while fitting the next proposal.

The adaptive Gaussian updates

For $Z_i \stackrel{\text{iid}}{\sim} q_k$ and $\alpha_i = f(Z_i)p(Z_i)/q_k(Z_i)$,

$$\mu_{k+1,j} = \frac{\sum_i \alpha_i Z_{ij}}{\sum_i \alpha_i},$$

$$v_{k+1,j} = \frac{\sum_i \alpha_i (Z_{ij} - \mu_{k+1,j})^2}{\sum_i \alpha_i}.$$

- Compute the new mean first.
- Centre the variance around that new mean.
- Evaluate $f(Z_i)$ on the complete sample vector.
- Require positive total weight and address zero fitted variances.