Trust Regions, ANEMONE, and ML Jargon Intuition

Anabel Yong

22nd April 2026

Abstract

A first-order Taylor approximation means replacing the true log-likelihood by its tangent-line approximation around the current parameters. Full-batch EM for PCs can then be seen as maximizing this local linearized objective while paying a KL penalty for changing the current joint distribution $p_\phi(X, Z)$ too much. The KL term is therefore a trust region in distribution space: it lets the model improve likelihood but only through a distributional step that remains close to the current model.

1 What We Need to Know about Probabilistic Circuits

A probabilistic circuit (PC) is a distribution built from sum nodes (mixtures), product nodes (factorizations) and input nodes (simple univariate distributions). Formally, a sum node computes a weighted mixture of its children, and a product node multiplies child distributions. The edge weights $\theta_{n,c}$ on sum nodes are learnable parameters. This paper [1] uses log-parameters $\phi_{n,c} := \log \theta_{n,c}$. PCs are viewed as latent-variable models: each sum node introduces a discrete latent variable that chooses one child.

2 Full-batch Expectation Maximisation (EM)

EM is used when the model has hidden variables Z. Instead of maximizing:

$$\log p_\phi(x) = \log \sum_z p_\phi(x, z)$$

directly, EM maximizes the surrogate:

$$\mathcal{Q}_\phi(\phi') = \sum_z p_\phi(z|x) \log p_{\phi'}(x, z)$$

For a dataset D, Liu, Shao and Broeck [1] uses the average over all samples. Full batch here means **use the whole dataset before making one parameter update**. For PCs, this leads to the standard flow/count update:

$$\theta'_{n,c} = \frac{F^D_\phi(n, c)}{Z_n}$$

where $F^D_\phi(n, c)$ is the average flow through edge (n, c), and Z_n normalizes the outgoing weights of sum node n.

3 First-order Taylor approximation of Log-likelihood

This phrase to me, sounds scary, but it is just the tangent-line idea from calculus. If we have a function $f(\phi)$, then near the current point ϕ:

$$f(\phi') \approx f(\phi) + \langle \nabla f(\phi), \phi' - \phi \rangle.$$

So for log-likelihood, around the current parameters ϕ, we replace the real objective by its **local linear approximation**:

$$\log p_{\phi'}(x) \approx \log p_\phi(x) + \left\langle \frac{\delta \log p_\phi(x)}{\delta \phi}, \phi' - \phi \right\rangle$$

This means the constant term $\log p_\phi(x)$ is where we currently are. The gradient says which direction increases likelihood fastest nearby. The approximation is only trustworthy locally, not for huge moves. This is what the first order Taylor approximation means.

4 Where does KL trust region come in?

The paper [1]'s Proposition 1, rewrites the full-batch EM objective as, up to constants:

$$\mathcal{Q}_{\phi}^{D}(\phi') = \frac{1}{|D|}\sum_{x\in D}\left\langle \frac{\delta \log p_{\phi}(x)}{\delta \phi}.\phi' \right\rangle - \mathbf{KL}_{\phi}(\phi'),$$

where:

$$\mathbf{KL}_{\phi}(\phi') := D_{KL}(p_{\phi}(X,Z)||p_{\phi'}(X,Z))$$

. Then.

$$\frac{1}{|D|}\sum_{x\in D}\left[\log p_{\phi}(x) + \left\langle \frac{\delta \log p_{\phi}(x)}{\delta \phi}.\phi' - \phi \right\rangle\right] - \mathbf{KL}_{\phi}(\phi')$$

This means that EM is maximizing a linearized log likelihood improvement term and minus a KL penalty that discourages the new distribution from drifting too far from the current one. The KL term is a trust region in a distribution space. The update says improve likelihood but only by making a distributional move I trust. Not "do not change parameters too much", but "do not change the actual probability distribution too much". Two parameter vectors can be far apart numerically while representing nearly the same distribution and vice versa. However, Liu, Shao and Broeck [1] explicitly contrasts this with gradient methods using an L_2 penalty in parameter space, arguing that KL is the more natural geometry because it measures closeness of distributions, not raw parameter coordinates. So, this is useful as it explains why **full-batch EM is often stable**. A pure gradient-style linearization would say: "follow the slope". But a linear approximation can be wildly wrong if you move too far. The KL term keeps the step conservative in the right space: the space of distributions. So EM is not merely greedy. It is greedy but regularized.

5 Returning to PCs

In PCs, parameters sit on sum-node edges and must remain locally normalized:

$$\sum_{c\in ch(n)} \theta_{n,c} = 1 \qquad \text{for every sum node } n$$

The EM update automatically respects these normalization constraints, which is one reason it is so natural for PCs. The paper shows the full batch update emerges after combining the objective with these normalization constants.

5.1 What is the Trust Region idea, without jargon?

A trust region just means **"I trust my local approximation only within a neighbourhood around the current model."** Here the neighbourhood is defined by KL divergence:

$$D_{KL}(p_{\phi}||p_{\phi'}) \leq \varepsilon$$

The gradient here says where to go, and the KL says how radically the probability distribution is allowed to change. So the trust region in distribution space does not mean "small Euclidean move in ϕ". It means "small change in what probabilities the model assigns".

Example intuition. Consider a PC whose root sum node mixes two components:

$$p_{\theta}(x) = \theta_1 p_1(x) + \theta_2 p_2(x), \qquad \theta_1 + \theta_2 = 1.$$

Suppose the current parameters are $\theta_1 = \theta_2 = 0.5$. If the local data evidence favors the first component, the first-order term encourages increasing θ_1. However, the KL regularizer prevents the update from changing the whole distribution too abruptly. Thus, instead of making an overly aggressive move that overfits the current evidence, EM chooses a conservative update that improves likelihood while keeping the new model distribution close to the old one. This is precisely the sense in which EM performs a trust-region step in distribution space. The paper further uses this viewpoint to motivate stronger KL regularization in the mini-batch setting, where limited data can otherwise induce unstable updates.

6 Why Full-Batch EM is special.

Because with the whole dataset, the linearized likelihood term already reflects the global data distribution. So the KL regularizer is enough to make a principled stable update. In the minibatch case, the batch only gives a noisy partial view of teh data. The paper [1]'s idea is: if the evidence is less reliable, increase the weight on the KL terms. That yields their update:

$$\theta'_{n,c} - \frac{TD_\phi(n)\theta_{n,c} + \eta F^D_\phi(n,c)}{Z_n}$$

which blends the current parameter with the batch evidence in a way controlled by top-down importance $TD_\phi(n)$. So full-batch EM is the clean version of the idea, minibatch comes from strenghtening the trust region when the data signal is noisier.

References

1. Liu A, Shao Z and Broeck GV den. Rethinking Probabilistic Circuit Parameter Learning. 2025. arXiv: 2505.19982 [cs.LG]. Available from: https://arxiv.org/abs/2505.19982