Data Modelling Methods-VI

CS771: Introduction to Machine Learning

Purushottam Kar

Outline of today's discussion

- Two alternating optimization algorithms for PPCA
- The EM algorithm

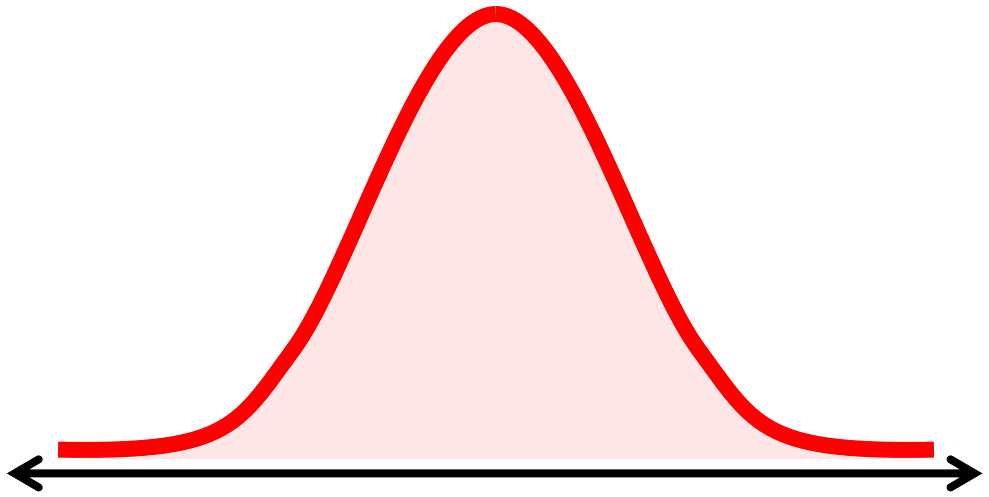
- Please respond to the Piazza poll on lecture topics
- This directly affects the content of the course
- So far very few have responded

Alternating Optimization for PCA

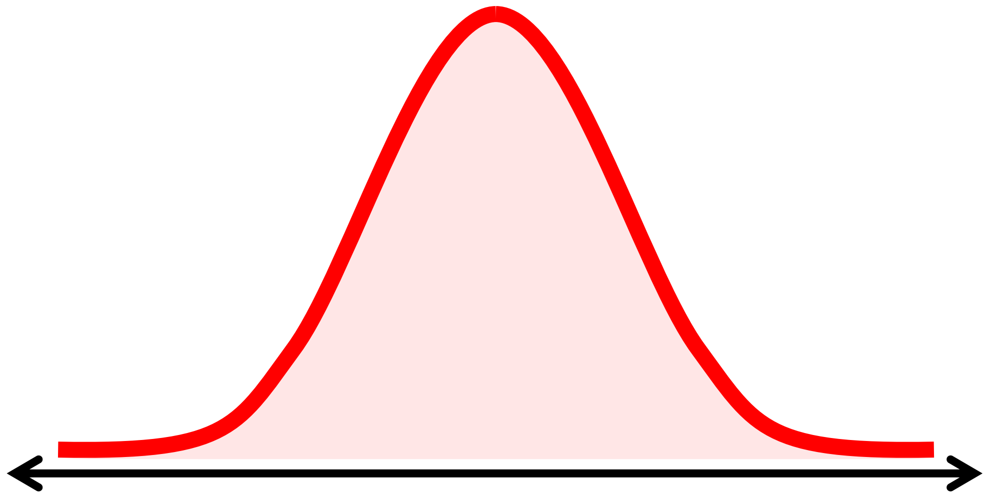
Low-dimensional Structure in Data

Low-dimensional Structure in Data

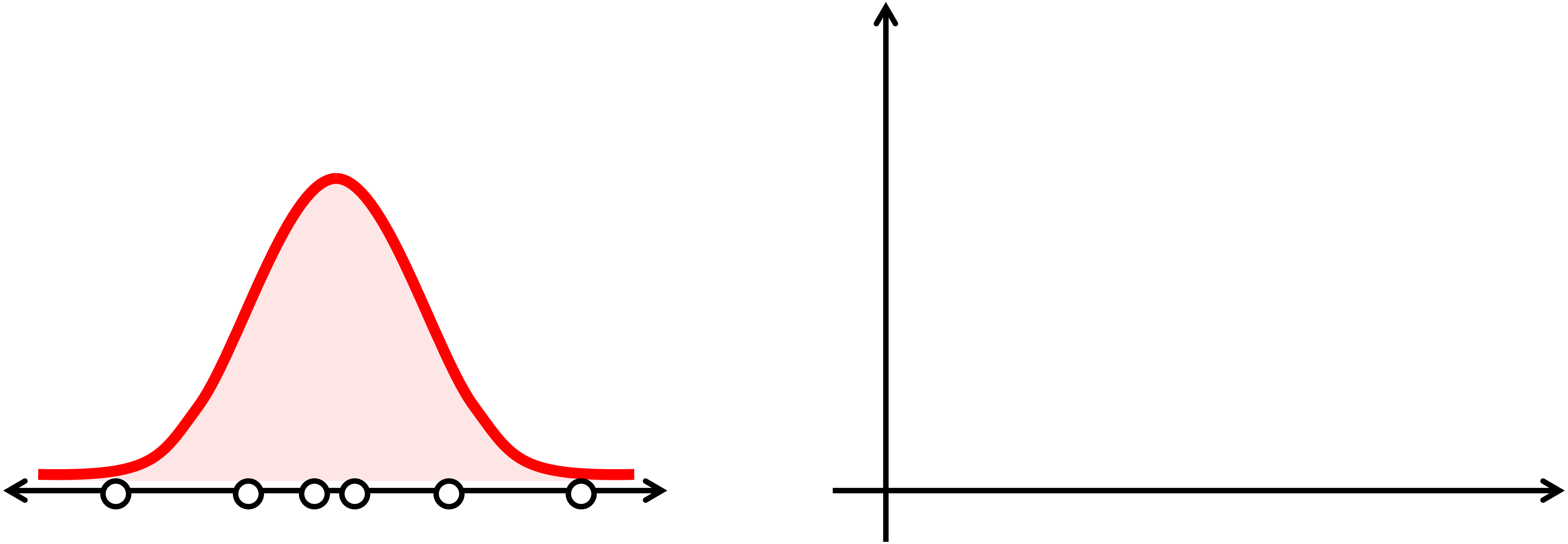
Low-dimensional Structure in Data



Low-dimensional Structure in Data

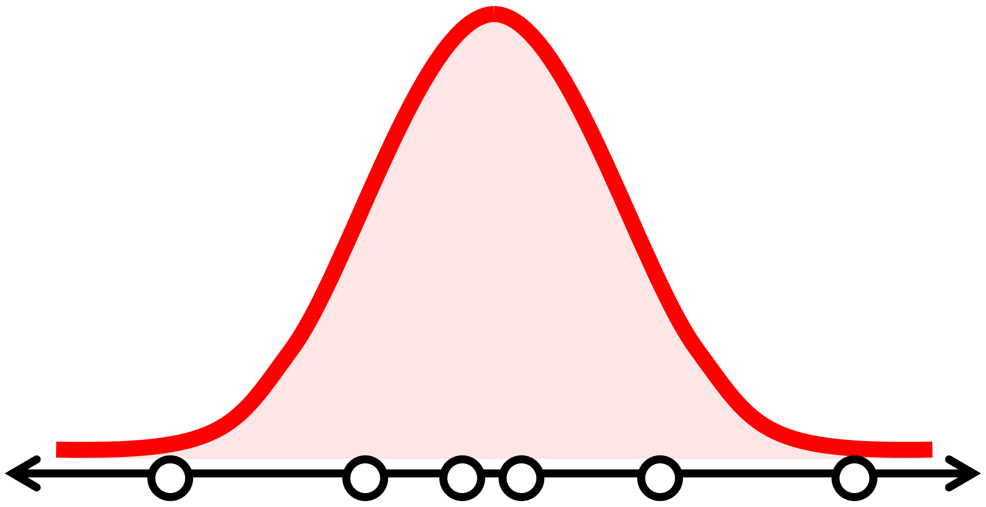


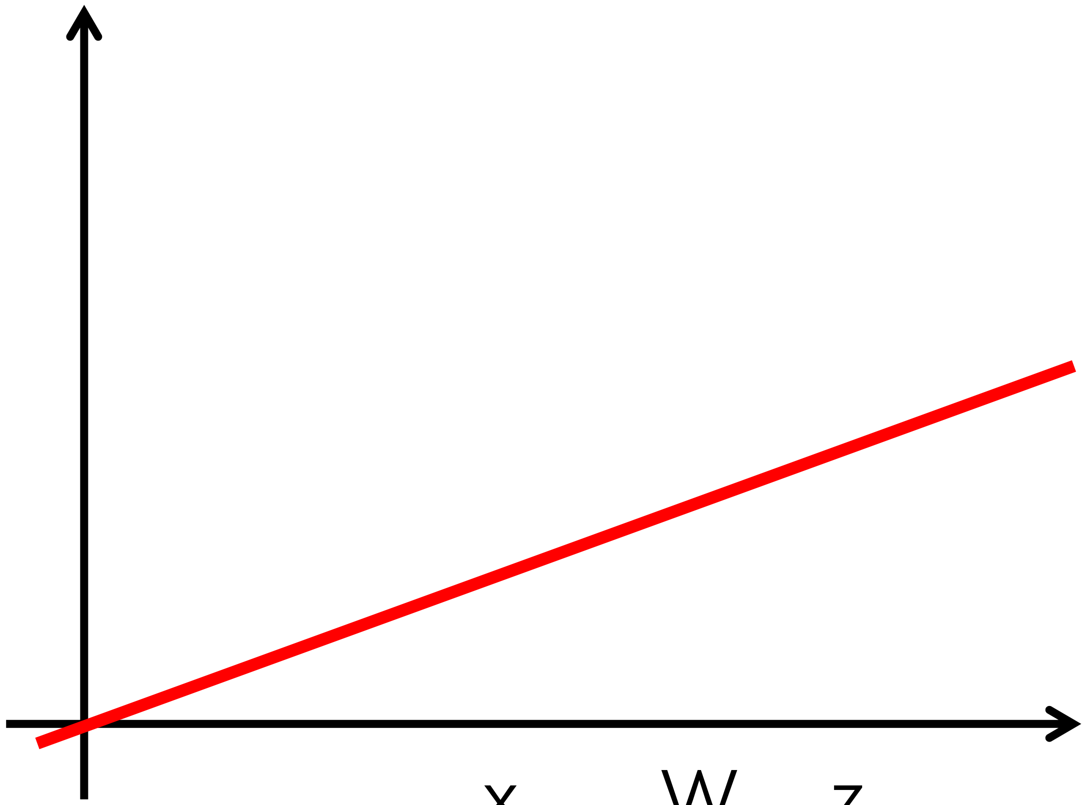
Low-dimensional Structure in Data



$$\mathbf{z}^i \sim \mathcal{N}(\mathbf{0}, I_k)$$

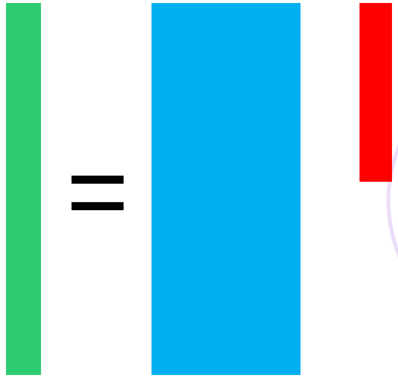
Low-dimensional Structure in Data



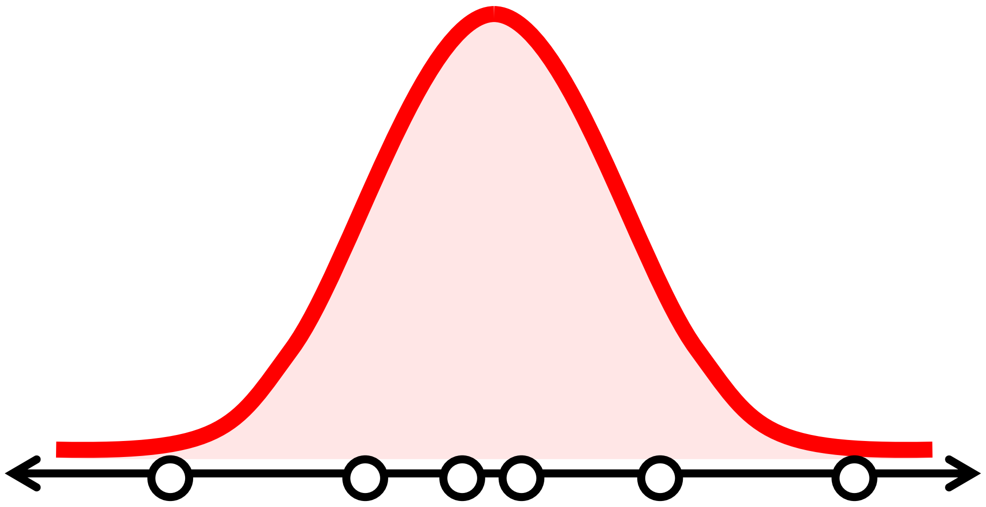


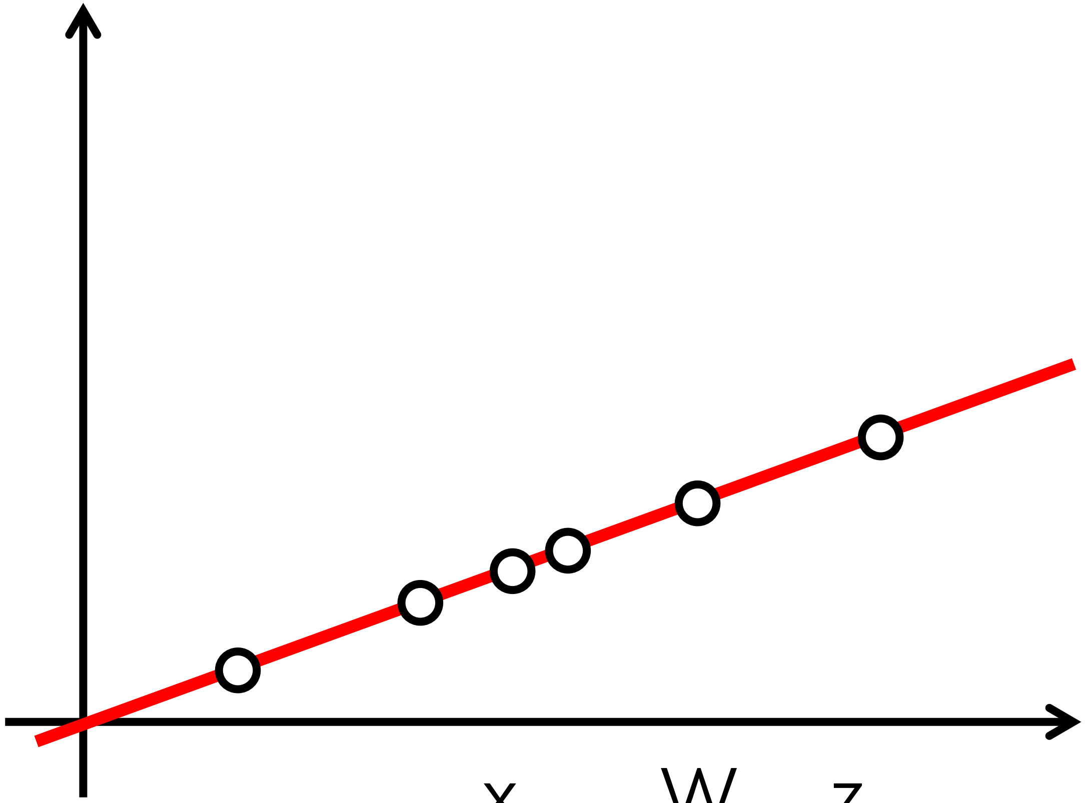
$$\mathbf{z}^i \sim \mathcal{N}(\mathbf{0}, I_k) \qquad \mathbf{x}^i = W\mathbf{z}^i$$

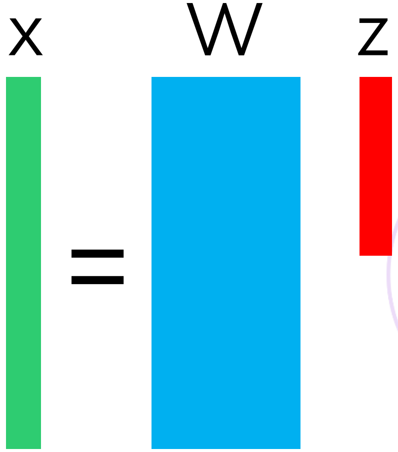
$$W \in \mathbb{R}^{d \times k}$$



Low-dimensional Structure in Data



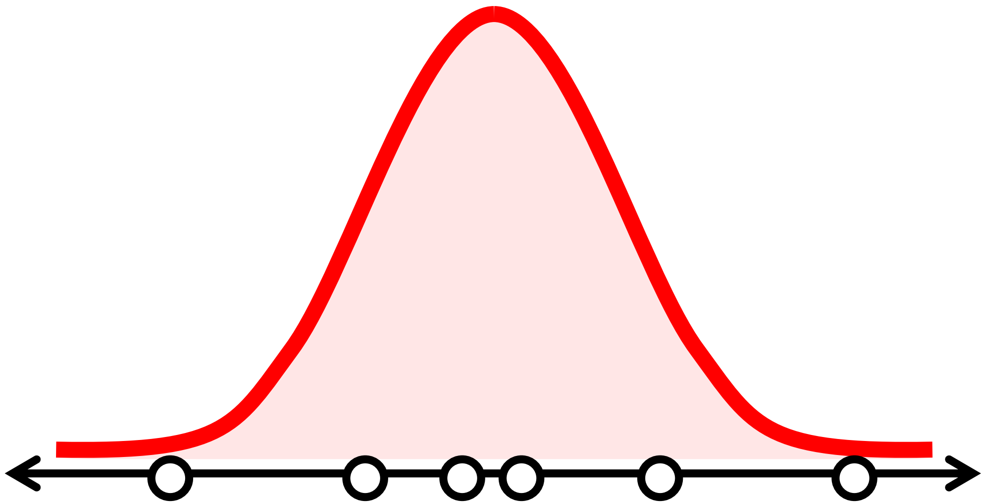


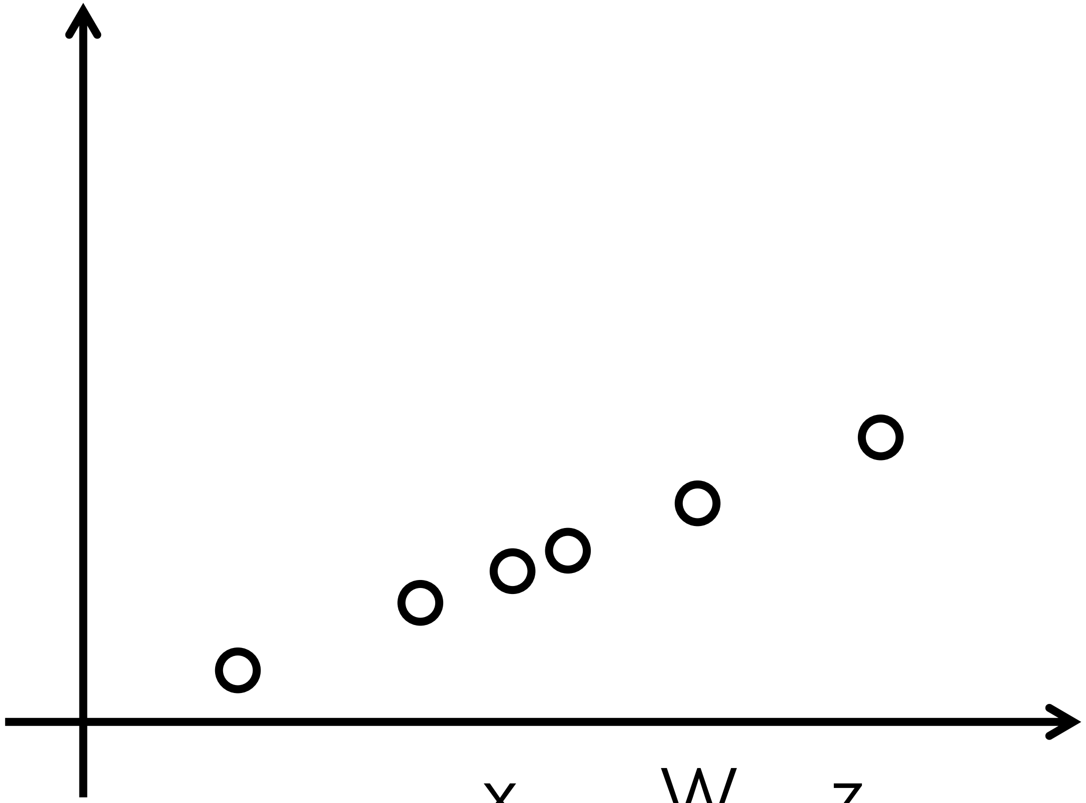


$$\mathbf{z}^i \sim \mathcal{N}(\mathbf{0}, I_k) \qquad \mathbf{x}^i = W\mathbf{z}^i$$

$$W \in \mathbb{R}^{d \times k}$$

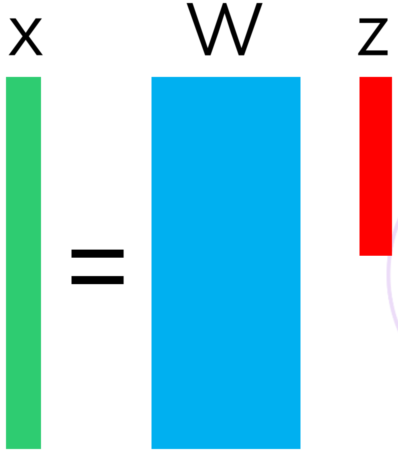
Low-dimensional Structure in Data



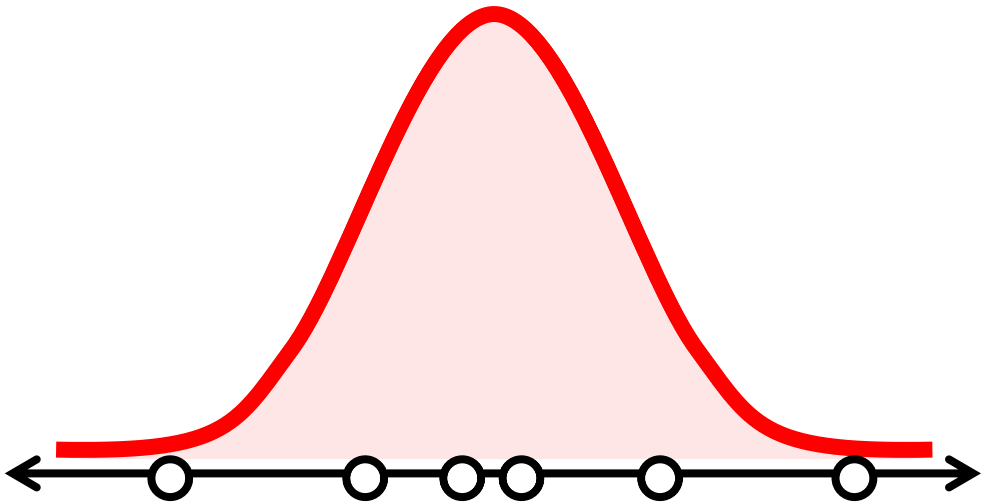


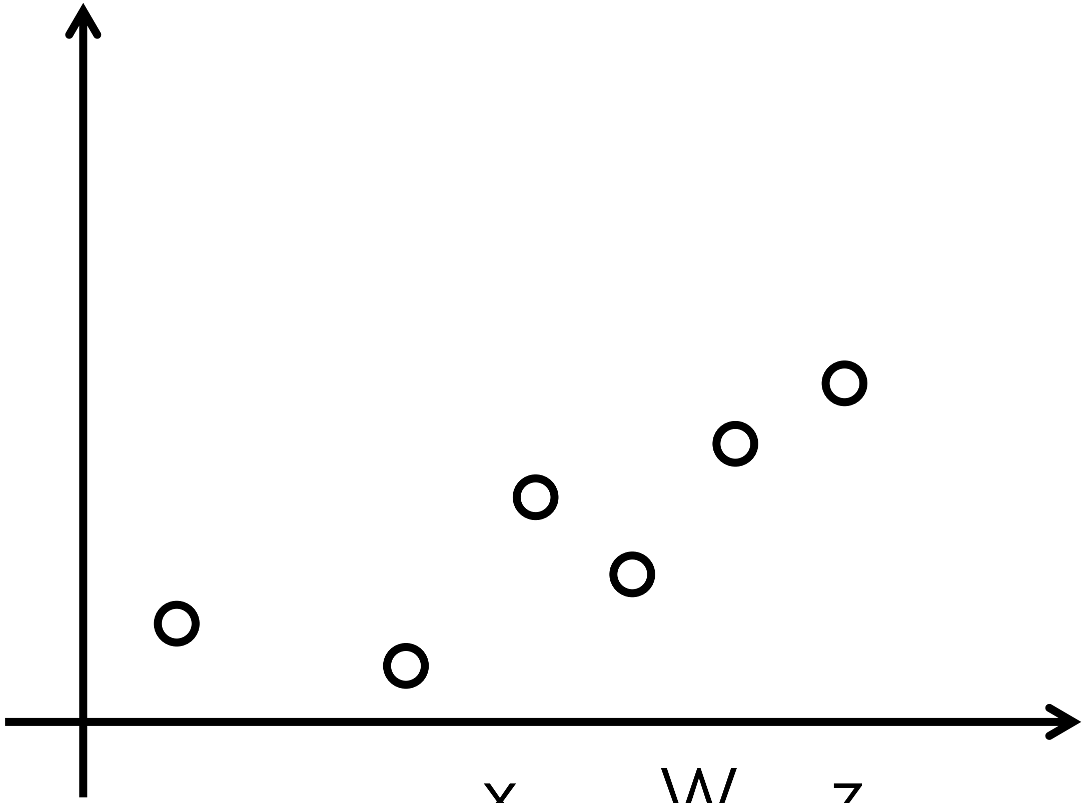
$$\mathbf{z}^i \sim \mathcal{N}(\mathbf{0}, I_k) \qquad \mathbf{x}^i = W\mathbf{z}^i$$

$$W \in \mathbb{R}^{d \times k}$$



Low-dimensional Structure in Data

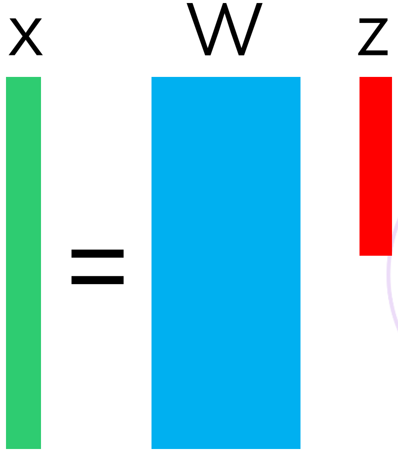




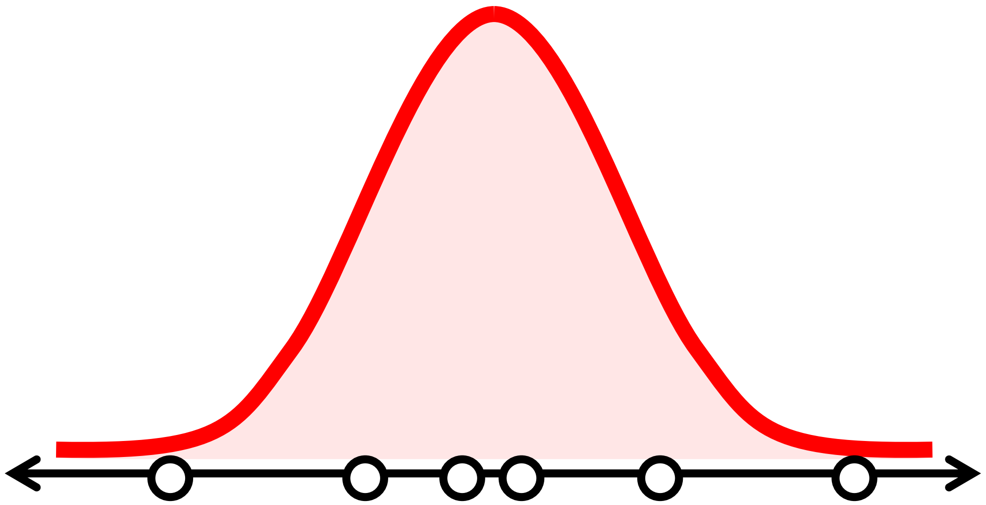
$$\mathbf{z}^i \sim \mathcal{N}(\mathbf{0}, I_k)$$

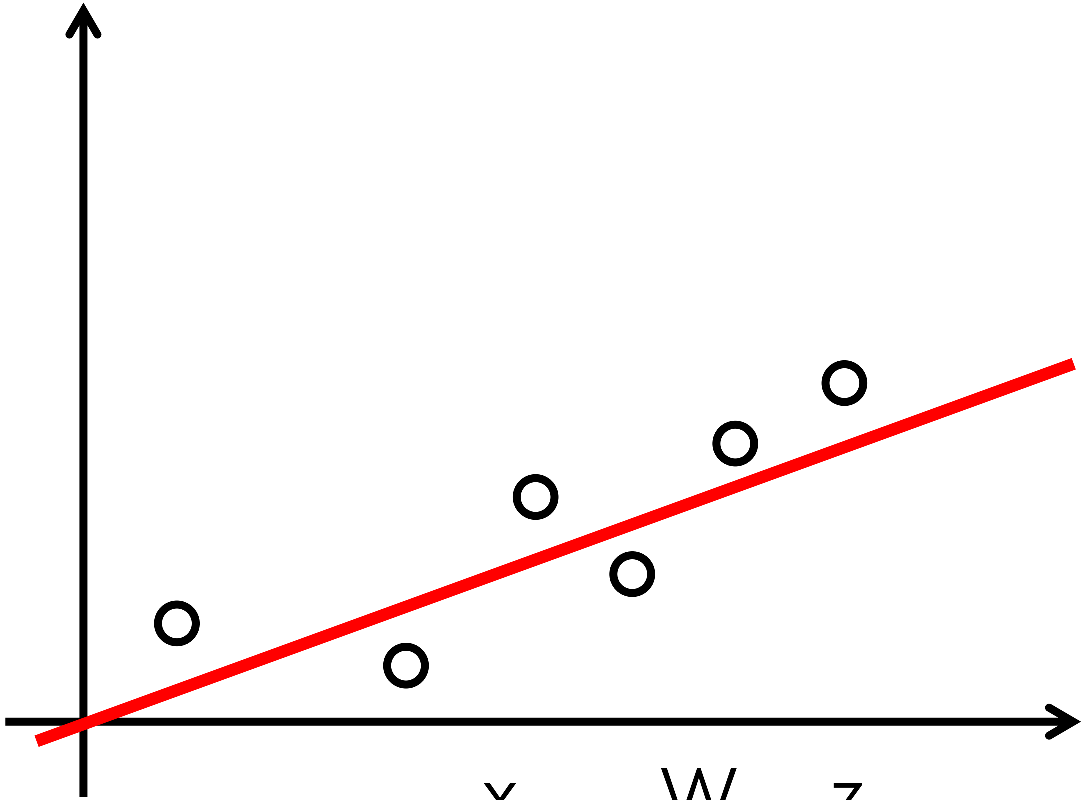
$$\mathbf{x}^i = W\mathbf{z}^i + \boldsymbol{\epsilon}^i, \;\; \boldsymbol{\epsilon}^i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \cdot I_d)$$

$$W \in \mathbb{R}^{d \times k}$$



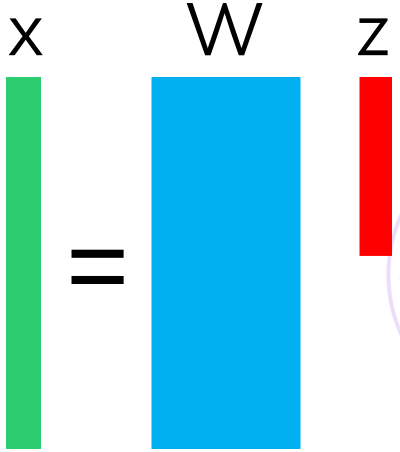
Low-dimensional Structure in Data



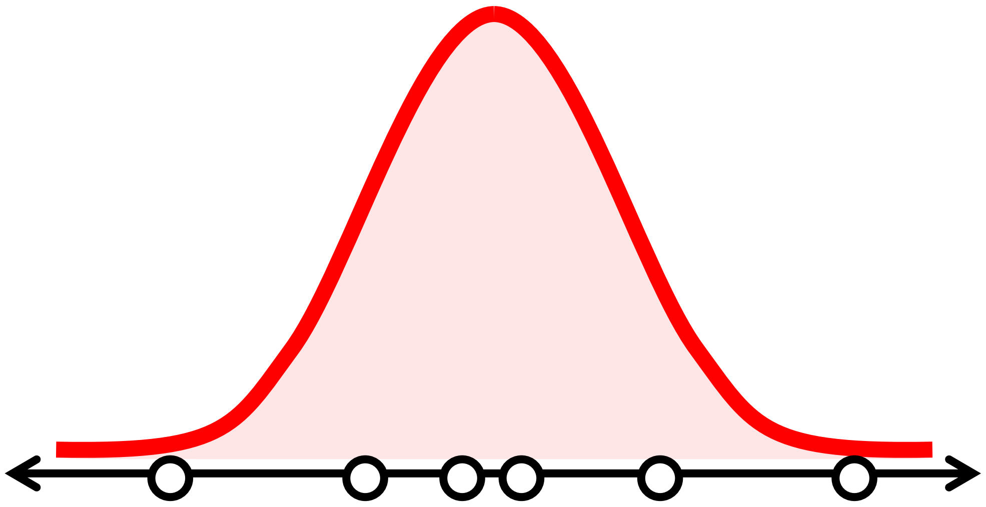


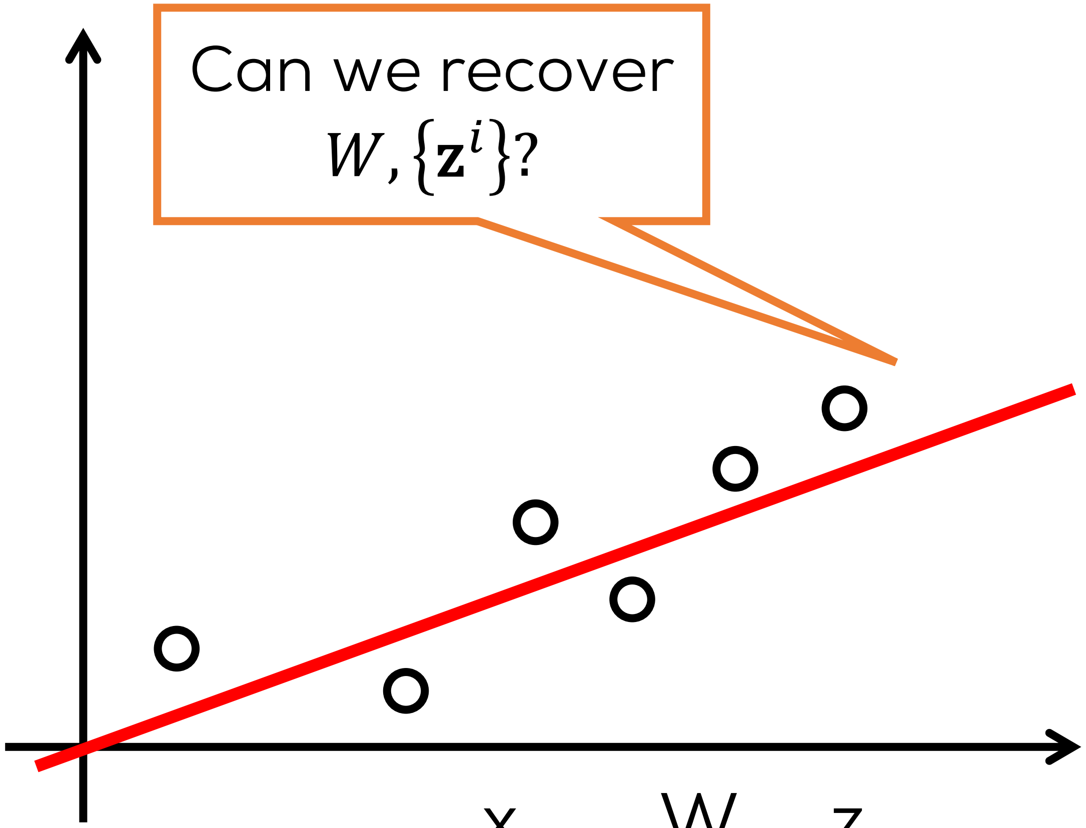
$$\mathbf{z}^i \sim \mathcal{N}(\mathbf{0}, I_k) \qquad \mathbf{x}^i = W\mathbf{z}^i + \boldsymbol{\epsilon}^i,\ \ \boldsymbol{\epsilon}^i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \cdot I_d)$$

$$W \in \mathbb{R}^{d \times k}$$



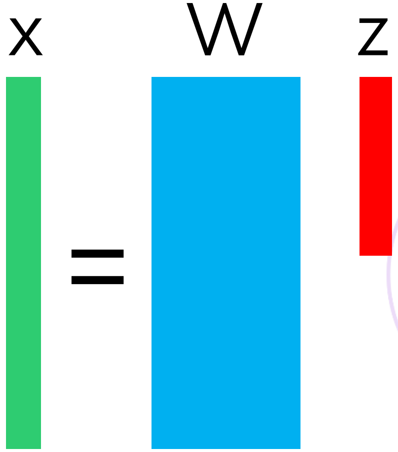
Low-dimensional Structure in Data



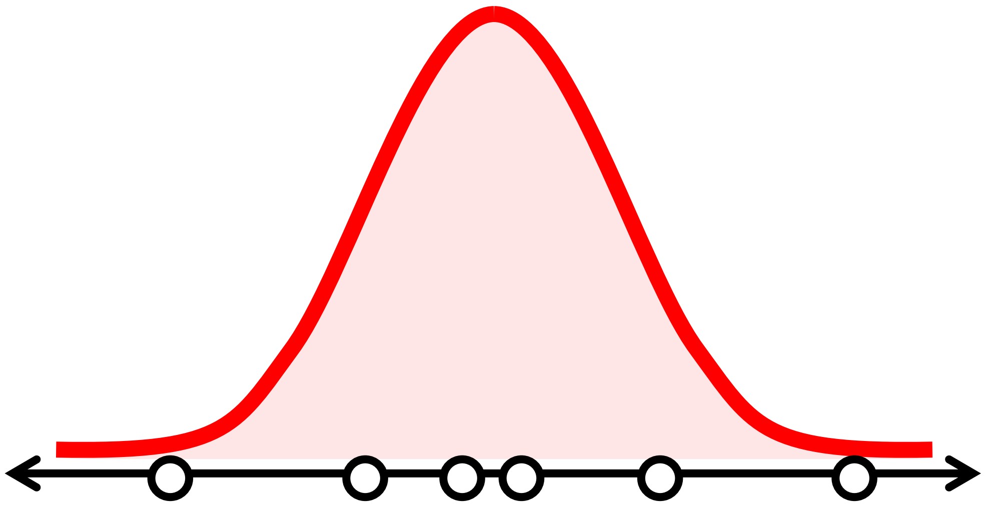


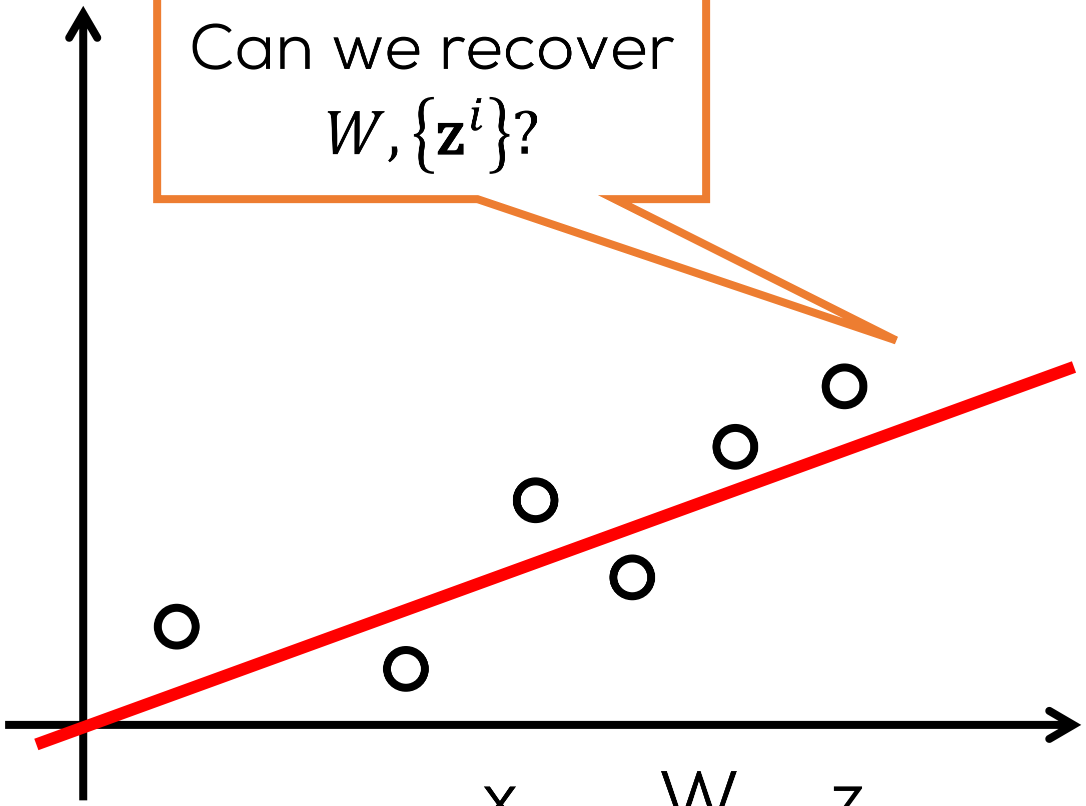
$$\mathbf{z}^i \sim \mathcal{N}(\mathbf{0}, I_k) \qquad \mathbf{x}^i = W\mathbf{z}^i + \boldsymbol{\epsilon}^i,\ \boldsymbol{\epsilon}^i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \cdot I_d)$$

$$W \in \mathbb{R}^{d \times k}$$



Low-dimensional Structure in Data





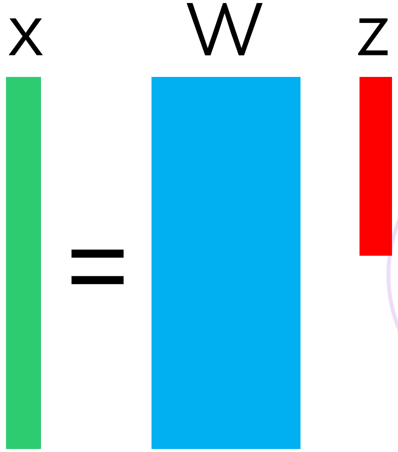
Can we recover $W, \{\mathbf{z}^i\}$?

$\mathbf{z}^i \sim \mathcal{N}(\mathbf{0}, I_k)$

$\mathbf{x}^i = W\mathbf{z}^i + \boldsymbol{\epsilon}^i, \ \boldsymbol{\epsilon}^i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \cdot I_d)$

Dictionary/Factor Loading matrix

$W \in \mathbb{R}^{d \times k}$



Marginals and Posteriors

- We have $\mathbf{z}^i \sim \mathcal{N}(\mathbf{0}, I_k)$ and $\mathbf{x}^i|\mathbf{z}^i \sim \mathcal{N}\left(W\mathbf{z}^i, \sigma^2 \cdot I_d\right)$

- **Marginal Distribution**
  $\mathbb{P}\left[\mathbf{x}^i \mid \sigma, W\right] = \mathcal{N}(\mathbf{0}, \Sigma_x)$
  where $\Sigma_x = WW^\top + \sigma^2 \cdot I_d \in \mathbb{R}^{d\times d}$

- **Posterior Distribution**
  $\mathbb{P}\left[\mathbf{z}^i \mid \mathbf{x}^i, \sigma, W\right] = \mathcal{N}\left(\boldsymbol{\mu}_z^i, \Sigma_z\right)$
  where $\boldsymbol{\mu}_z^i = (W^\top W + \sigma^2 \cdot I_k)^{-1} W^\top \mathbf{x}^i \in \mathbb{R}^k$
  and $\Sigma_z = \sigma^2 \cdot (W^\top W + \sigma^2 \cdot I_k)^{-1} \in \mathbb{R}^{k\times k}$

Marginals and Posteriors

- We have $\mathbf{z}^i \sim \mathcal{N}(\mathbf{0}, I_k)$ and $\mathbf{x}^i | \mathbf{z}^i \sim \mathcal{N}\left(W\mathbf{z}^i, \sigma^2 \cdot I_d\right)$

Suppose $\left\|\mathbf{x}^i\right\|_2$ is large. It is unlikely that noise cause it. More likely that $\mathbf{z}^i \neq \mathbf{0}$

Because we have seen $\mathbf{x}^i$

- **Marginal Distribution**
  $\mathbb{P}\left[\mathbf{x}^i \mid \sigma, W\right] = \mathcal{N}(\mathbf{0}, \Sigma_x)$
  where $\Sigma_x = WW^\top + \sigma^2 \cdot I_d \in \mathbb{R}^{d\times d}$

Why isn't $\boldsymbol{\mu}_z^i = \mathbf{0}$?

- **Posterior Distribution**
  $\mathbb{P}\left[\mathbf{z}^i \mid \mathbf{x}^i, \sigma, W\right] = \mathcal{N}\left(\boldsymbol{\mu}_z^i, \Sigma_z\right)$
  where $\boldsymbol{\mu}_z^i = (W^\top W + \sigma^2 \cdot I_k)^{-1} W^\top \mathbf{x}^i \in \mathbb{R}^k$
  and $\Sigma_z = \sigma^2 \cdot (W^\top W + \sigma^2 \cdot I_k)^{-1} \in \mathbb{R}^{k\times k}$

Marginals and Posteriors

- We have $\mathbf{z}^i \sim \mathcal{N}(\mathbf{0}, I_k)$ and $\mathbf{x}^i | \mathbf{z}^i \sim \mathcal{N}\left(W\mathbf{z}^i, \sigma^2 \cdot I_d\right)$

- **Marginal Distribution**
  $\mathbb{P}\left[\mathbf{x}^i \mid \sigma, W\right] = \mathcal{N}(\mathbf{0}, \Sigma_x)$
  where $\Sigma_x = WW^\top + \sigma^2 \cdot I_d \in \mathbb{R}^{d \times d}$

- **Posterior Distribution**
  $\mathbb{P}\left[\mathbf{z}^i \mid \mathbf{x}^i, \sigma, W\right] = \mathcal{N}\left(\boldsymbol{\mu}_z^i, \Sigma_z\right)$
  where $\boldsymbol{\mu}_z^i = (W^\top W + \sigma^2 \cdot I_k)^{-1} W^\top \mathbf{x}^i \in \mathbb{R}^k$
  and $\Sigma_z = \sigma^2 \cdot (W^\top W + \sigma^2 \cdot I_k)^{-1} \in \mathbb{R}^{k \times k}$

Suppose $\|\mathbf{x}^i\|_2$ is large. It is unlikely that noise cause it. More likely that $\mathbf{z}^i \neq \mathbf{0}$

Because we have seen $\mathbf{x}^i$

Why isn't $\boldsymbol{\mu}_z^i = \mathbf{0}$?

Because $\mathbf{x}^i = W\mathbf{z}^i + \boldsymbol{\epsilon}^i$, where $\boldsymbol{\epsilon}^i \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \cdot I_d\right)$ and least squares is the MLE

For Gaussians, mean is mode

Why does this look like least squares?

[**BIS**] eqn (12.42) has a mistake

Estimating $\mathbf{z}^i$

- Recall, in the last lecture, we claimed that in PML setting, $W_{\text{MLE}} = U_k\sqrt{\Lambda_k} \in \mathbb{R}^{d\times k}$ and $\mathbf{z}^i = \Lambda_k^{-1} W_{\text{MLE}}^\top \mathbf{x}^i$
- We can easily derive this now as the MLE estimate for $\mathbf{z}^i$
- We have

$$\underset{\mathbf{z}^i}{\operatorname{argmax}}\ \mathbb{P}\left[\mathbf{z}^i \mid \mathbf{x}^i, \sigma, W\right] = \boldsymbol{\mu}_z^i$$

$\mathbb{P}\left[\mathbf{z}^i \mid \mathbf{x}^i, \sigma, W\right] = \mathcal{N}\left(\boldsymbol{\mu}_z^i, \Sigma_z\right)$
mode of a Gaussian is mean

$$= (W^\top W + \sigma^2 \cdot I_k)^{-1} W^\top \mathbf{x}^i$$

See previous slide

$$= (\Lambda_k + \sigma^2 \cdot I_k)^{-1} W^\top \mathbf{x}^i$$

$U_k^\top U_k = I_k$

$$= \Lambda_k^{-1} W^\top \mathbf{x}^i$$

If $\sigma = 0$

- Note that we also have $\mathbb{E}\left[\mathbf{z}^i \mid \mathbf{x}^i, \sigma, W\right] = (\Lambda_k + \sigma^2 \cdot I_k)^{-1} W^\top \mathbf{x}^i$

The Complete Likelihood

- Given data $X = [\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n] \in \mathbb{R}^{d \times n}$
- Let (latent) low-dim reps be $Z = [\mathbf{z}^1, \mathbf{z}^2, \ldots, \mathbf{z}^n] \in \mathbb{R}^{k \times n}$
- Observed data likelihood $\mathbb{P}[X \mid W, \sigma] = \prod_{i=1}^n \mathbb{P}\left[\mathbf{x}^i \mid W, \sigma\right]$
- Complete data likelihood $\mathbb{P}[X, Z \mid W, \sigma] = \prod_{i=1}^n \mathbb{P}\left[\mathbf{x}^i, \mathbf{z}^i \mid W, \sigma\right]$

$$\log \mathbb{P}[X, Z \mid W, \sigma] = \sum_{i=1}^n \log \mathbb{P}\left[\mathbf{x}^i \mid \mathbf{z}^i, W, \sigma\right] + \log \mathbb{P}\left[\mathbf{z}^i \mid W, \sigma\right]$$

$$= -\sum_{i=1}^n \frac{1}{2\sigma^2}\left(\left\|\mathbf{x}^i\right\|_2^2 + \left(\mathbf{z}^i\right)^\top W^\top W \mathbf{z}^i - 2\left(\mathbf{z}^i\right)^\top W^\top \mathbf{x}^i\right) + \frac{1}{2} \cdot \left\|\mathbf{z}^i\right\|_2^2 + C$$

$$= -\sum_{i=1}^n \frac{1}{2\sigma^2}\left(\mathrm{tr}\left(W^\top W \mathbf{z}^i \left(\mathbf{z}^i\right)^\top\right) - 2\left(\mathbf{z}^i\right)^\top W^\top \mathbf{x}^i\right) + \frac{1}{2} \cdot \mathrm{tr}\left(\mathbf{z}^i \left(\mathbf{z}^i\right)^\top\right) + C'$$

The Complete Likelihood

- Given data $X = [\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n] \in \mathbb{R}^{d\times n}$
- Let (latent) low-dim reps be $Z = [\mathbf{z}^1, \mathbf{z}^2, \dots, \mathbf{z}^n] \in \mathbb{R}^{k\times n}$
- We have

$$\arg\max_W \log \mathbb{P}[X, Z \mid W, \sigma]$$

$$= \arg\min_W \sum_{i=1}^n \frac{1}{\sigma^2}\left(\mathrm{tr}\left(W^\top W \mathbf{z}^i (\mathbf{z}^i)^\top\right) - 2(\mathbf{z}^i)^\top W^\top \mathbf{x}^i\right) + \mathrm{tr}\left(\mathbf{z}^i (\mathbf{z}^i)^\top\right)$$

$$= \arg\min_W \sum_{i=1}^n \mathrm{tr}\left(W^\top W \mathbf{z}^i (\mathbf{z}^i)^\top\right) - 2(\mathbf{z}^i)^\top W^\top \mathbf{x}^i$$

$$= \left[\sum_{i=1}^n \mathbf{x}^i (\mathbf{z}^i)^\top\right] \cdot \left[\sum_{i=1}^n \mathbf{z}^i (\mathbf{z}^i)^\top\right]^{-1}$$

- For simplicity, assume σ is known (can estimate it too)

The Complete Likelihood

- Given data $X = [\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n] \in \mathbb{R}^{d \times n}$
- Let (latent) low-dim reps be $Z = [\mathbf{z}^1, \mathbf{z}^2, \ldots, \mathbf{z}^n]$
- We have

$$\arg\max_W \log \mathbb{P}[X, Z \mid W, \sigma]$$

$$= \arg\min_W \sum_{i=1}^n \frac{1}{\sigma^2}\left(\mathrm{tr}\left(W^\top W \mathbf{z}^i (\mathbf{z}^i)^\top\right) - 2(\mathbf{z}^i)^\top W^\top \mathbf{x}^i\right) + \mathrm{tr}\left(\mathbf{z}^i (\mathbf{z}^i)^\top\right)$$

$$= \arg\min_W \sum_{i=1}^n \mathrm{tr}\left(W^\top W \mathbf{z}^i (\mathbf{z}^i)^\top\right) - 2(\mathbf{z}^i)^\top W^\top \mathbf{x}^i$$

$$= \left[\sum_{i=1}^n \mathbf{x}^i (\mathbf{z}^i)^\top\right] \cdot \left[\sum_{i=1}^n \mathbf{z}^i (\mathbf{z}^i)^\top\right]^{-1}$$

- For simplicity, assume σ is known (can estimate it too)

Actually doing least squares to minimize reconstruction error ☺

$$\arg\min_W \sum_{i=1}^n \left\|\mathbf{x}^i - W\mathbf{z}^i\right\|_2^2$$

Apply first order optimality condition

Alternating Optimization

- So if someone gave me $\mathbf{z}^i$, I can estimate W as

$$W_{\mathrm{MLE}} = \left[\sum_i^n \mathbf{x}^i \left(\mathbf{z}^i\right)^\top\right] \cdot \left[\sum_i^n \mathbf{z}^i \left(\mathbf{z}^i\right)^\top\right]^{-1}$$

- However, if someone gave me W, I can estimate $\mathbf{z}^i$ as

$$\mathbf{z}^i_{\mathrm{MLE}} = (W^\top W + \sigma^2 \cdot I_k)^{-1} W^\top \mathbf{x}^i$$

- So can I do alternating optimization using these?
- Yes, of course!

Hard Alternating Minimization

HARD ALTERNATING OPTIMIZATION

1. Initialize W^0
2. For $t = 0, 1, 2, \ldots$
   1. For $i \in [n]$, update $\mathbf{z}^{i,t}$ using W^t
      1. Let $\mathbf{z}^{i,t} = \arg\max_{\mathbf{z}^i} \mathbb{P}\left[\mathbf{z}^i \mid \mathbf{x}^i, \sigma, W^t\right]$

         $= \left((W^t)^\top W^t + \sigma^2 \cdot I_k\right)^{-1} (W^t)^\top \mathbf{x}^i$

   2. Update $W^{t+1} = \arg\max_{W} \mathbb{P}\left[X, Z^t \mid W, \sigma\right]$

      $= \left[\sum_i^n \mathbf{x}^i (\mathbf{z}^{i,t})^\top\right] \cdot \left[\sum_i^n \mathbf{z}^{i,t} (\mathbf{z}^{i,t})^\top\right]^{-1}$

$O(k^2 d + dnk)$ time

$O(k^2 d + dnk)$ time

Hard Alternating Minimization for $\sigma = 0$

HARD ALTERNATING OPTIMIZATION

1. Initialize W^0
2. For $t = 0, 1, 2, \ldots$
   1. For $i \in [n]$, update $\mathbf{z}^{i,t}$ using W^t
      1. Let $\mathbf{z}^{i,t} = \arg\max_{\mathbf{z}^i} \mathbb{P}\left[\mathbf{z}^i \mid \mathbf{x}^i, W^t\right]$

$$= \left((W^t)^\top W^t\right)^{-1} (W^t)^\top \mathbf{x}^i$$

   2. Update $W^{t+1} = \arg\max_{W} \mathbb{P}\left[X, Z^t \mid W\right]$

$$= \left[\sum_i^n \mathbf{x}^i (\mathbf{z}^{i,t})^\top\right] \cdot \left[\sum_i^n \mathbf{z}^{i,t} (\mathbf{z}^{i,t})^\top\right]^{-1}$$

$O(k^2 d + dnk)$ time

$O(k^2 d + dnk)$ time

Some Thoughts

- We updated $\mathbf{z}^i$ using maximum posterior probability but W using maximum likelihood. Can we do MAP for W as well?
- Yes, put a prior on W but things get more complicated
- Can I have a "soft" assignment algorithm for this?
- Yes, but since there are infinite possibilities for $\mathbf{z}^i$, making partial assignments of a point $\mathbf{x}^i$ to all possible $\mathbf{z}^i$ is challenging
- Easier solution is to let $\mathbf{x}^i$ declare allegiance to expected values
- Use $\mathbb{E}\left[\mathbf{z}^i \mid \mathbf{x}^i, \sigma, W\right]$ and $\mathbb{E}\left[\mathbf{z}^i (\mathbf{z}^i)^\top \mid \mathbf{x}^i, \sigma, W\right]$ in AltOpt
- $\mathbb{E}\left[\mathbf{z}^i \mid \mathbf{x}^i, \sigma, W\right] = (W^\top W + \sigma^2 \cdot I_k)^{-1} W^\top \mathbf{x}^i = \boldsymbol{\zeta}^i$ (same as in hard algo)
- $\mathbb{E}\left[\mathbf{z}^i (\mathbf{z}^i)^\top \mid \mathbf{x}^i, \sigma, W\right] = \boldsymbol{\zeta}^i (\boldsymbol{\zeta}^i)^\top + \sigma^2 \cdot (W^\top W + \sigma^2 \cdot I_k)^{-1} = Z^i$

Some Thoughts

- We updated $\mathbf{z}^i$ using maximum posterior probability but W using maximum likelihood. Can we do MAP for W as well?
- Yes, put a prior on W but things get more complicated
- Can I have a "soft" assignment algorithm for this?
- Yes, but since there are infinite possibilities for $\mathbf{z}^i$, making partial assignments of a point $\mathbf{x}^i$ to all possible $\mathbf{z}^i$ is challenging
- Easier solution is to let $\mathbf{x}^i$ declare allegiance to expected
- Use $\mathbb{E}\left[\mathbf{z}^i \mid \mathbf{x}^i, \sigma, W\right]$ and $\mathbb{E}\left[\mathbf{z}^i (\mathbf{z}^i)^\top \mid \mathbf{x}^i, \sigma, W\right]$ in AltOpt
- $\mathbb{E}\left[\mathbf{z}^i \mid \mathbf{x}^i, \sigma, W\right] = (W^\top W + \sigma^2 \cdot I_k)^{-1} W^\top \mathbf{x}^i = \boldsymbol{\zeta}^i$ (same as in hard algo)
- $\mathbb{E}\left[\mathbf{z}^i (\mathbf{z}^i)^\top \mid \mathbf{x}^i, \sigma, W\right] = \boldsymbol{\zeta}^i (\boldsymbol{\zeta}^i)^\top + \sigma^2 \cdot (W^\top W + \sigma^2 \cdot I_k)^{-1} = Z^i$

Mode of Gaussian is the mean

This is new

Soft Alternating Minimization

SOFT ALTERNATING OPTIMIZATION

1. Initialize W^0
2. For $t = 0, 1, 2, \ldots$
   1. Compute $M^t = (W^t)^\top W^t + \sigma^2 \cdot I_k$
   2. For $i \in [n]$, update $\mathbf{z}^{i,t}$ and $Z^{i,t}$ using W^t
      1. Compute $\mathbf{z}^{i,t} = \mathbb{E}\left[\mathbf{z}^i \mid \mathbf{x}^i, \sigma, W^t\right] = (M^t)^{-1}(W^t)^\top \mathbf{x}^i$
      2. Compute $Z^{i,t} = \mathbb{E}\left[\mathbf{z}^i(\mathbf{z}^i)^\top \mid \mathbf{x}^i, \sigma, W\right] = \mathbf{z}^{i,t}\left(\mathbf{z}^{i,t}\right)^\top + \sigma^2 \cdot (M^t)^{-1}$
   3. Update $W^{t+1} = \arg\max_W \mathbb{E}[\log \mathbb{P}[X, Z^t \mid W, \sigma]]$

$$= \left[\sum_i^n \mathbf{x}^i\left(\mathbf{z}^{i,t}\right)^\top\right] \cdot \left[\sum_i^n Z^{i,t}\right]^{-1}$$

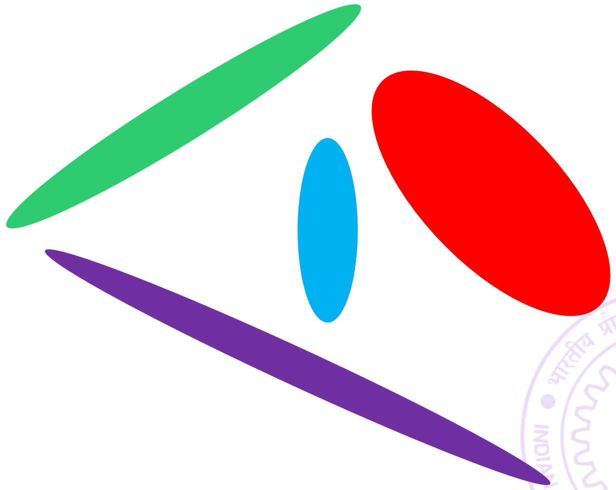
Same time complexity as hard AltOpt

Some Thoughts

- For $\sigma = 0$ hard and soft AltOpt are the same algorithm
- Power method-based solution took $O(dnk)$ time (last lecture)
- Most costly operation in power method: calculate $S\mathbf{v}, \mathbf{v} \in \mathbb{R}^d$

$$S\mathbf{v} = \frac{1}{n} \cdot XX^\top \mathbf{v}$$

- AltOpt solution takes $O(k^2 d + dnk)$ time – usually a bit more costly
- But … no convergence guarantee for AltOpt not even to the power-method solution ☹
- But … lots of things to play around with ☺
- Mixture of PPCA?
- Handle missing data $\mathbf{x} = [\mathbf{x}_{\text{obs}}, \mathbf{x}_{\text{miss}}]$
- Treat $\mathbf{x}_{\text{miss}}$ as latent vars and use $\mathbb{P}[\mathbf{x}] = \mathbb{P}[\mathbf{x}_{\text{miss}} \mid \mathbf{x}_{\text{obs}}] \cdot \mathbb{P}[\mathbf{x}_{\text{obs}}]$



The EM Algorithm

The EM Algorithm

- **EM**: Expectation Maximization
- Little secret: whenever we did "soft assignment" alternating optimization, we were executing exactly the EM algorithm
- Very versatile and adaptive to variety of problem settings
- Very popular for learning latent variable models
  soft k-means, HMMs (Baum-Welch), PPCA

The Generative Story with Latent Variables

- A parameterized distribution to generate two pairs of variables
$$\mathbb{P}[\mathbf{x}, \mathbf{z} \mid \boldsymbol{\Theta}^*], \mathbf{x} \in \mathcal{X}, \mathbf{z} \in \mathcal{Z}$$
- In secret a bunch of data points are generated …
$$(\mathbf{x}^1, \mathbf{z}^1), (\mathbf{x}^2, \mathbf{z}^2), \dots, (\mathbf{x}^n, \mathbf{z}^n)$$
- … but only the first component of each of them are revealed
$$\mathrm{X} = [\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n]$$
- Can we recover both $\boldsymbol{\Theta}^*$ as well as $\mathrm{Z} = [\mathbf{z}^1, \mathbf{z}^2, \dots, \mathbf{z}^n]$?
- **Clustering/GMM**: $\mathcal{Z} = \{1, 2, \dots, K\}$ for K clusters/Gaussians
- **Mixed Regression**: $\mathcal{Z} = \{0, 1\}$ for two components (may be more)
- **PCA/PPCA**: $\mathcal{Z} = \mathbb{R}^k$
- **Guess my Grocery List**:

MLE with Latent Variables

- We have

$$\mathbf{\Theta}_{\text{MLE}} = \arg\max_{\mathbf{\Theta}} \log \mathbb{P}[X \mid \mathbf{\Theta}] = \arg\max_{\mathbf{\Theta}} \sum_{i=1}^{n} \log \mathbb{P}\left[\mathbf{x}^i \mid \mathbf{\Theta}\right]$$

$$= \begin{cases} \arg\max_{\mathbf{\Theta}} \sum_{i=1}^{n} \log \sum_{\mathbf{z} \in \mathcal{Z}} \mathbb{P}\left[\mathbf{x}^i, \mathbf{z} \mid \mathbf{\Theta}\right], \text{when } \mathcal{Z} \text{ is discrete (GMM, MR)} \\ \arg\max_{\mathbf{\Theta}} \sum_{i=1}^{n} \log \int_{\mathbf{z} \in \mathcal{Z}} \mathbb{P}\left[\mathbf{x}^i, \mathbf{z} \mid \mathbf{\Theta}\right], \text{when } \mathcal{Z} \text{ is continuous (PCA)} \end{cases}$$

- Often NP-hard to solve these optimization problems directly
- Indirect methods required

Alternating Optimization to the Rescue

- In many of these problems, although $\max_{\boldsymbol{\Theta}} \log \mathbb{P}[X \mid \boldsymbol{\Theta}]$ is difficult,
- … but $\arg\max_{\boldsymbol{\Theta}} \log \mathbb{P}[X, Z \mid \boldsymbol{\Theta}]$ is simple
- … and $\arg\max_{Z} \log \mathbb{P}[Z \mid X, \boldsymbol{\Theta}]$ is simple
- Immediately leads us to a simple "hard" assignment algorithm

HARD ALTERNATING OPTIMIZATION

1. Initialize $\boldsymbol{\Theta}^0$
2. Update $Z^{t+1} = \arg\max_{Z} \mathbb{P}\left[Z \mid X, \boldsymbol{\Theta}\right]$
3. Update $\boldsymbol{\Theta}^{t+1} = \arg\max_{\boldsymbol{\Theta}} \mathbb{P}\left[X, Z \mid \boldsymbol{\Theta}\right]$
4. Repeat until convergence

A "Softer" Approach

- Hard AltOpt is often fast and used due to its speed
- However, for complicated problems it can misbehave
- Hard AltOpt also throws away useful information about the latent variables that don't win the MAP contest
- Motivates a softer approach which doesn't trust a single $\mathbf{z}^i$ for $\mathbf{x}^i$
- How to trust multiple latent variables?
  - Trust all of them but with varying degree of confidence?
  - Trust a few of them?
  - Trust a random latent variable (sampled from some distribution)?
- Is there a sound way to guide this choice?
- Yes, that way will give us the EM algorithm

Deriving the EM algorithm

- Assume discrete latent variables (for simplicity)
- Suppose we already have an estimate $\boldsymbol{\Theta}^0$ with us somehow

$$\log \mathbb{P}\left[\mathbf{x}^i \mid \boldsymbol{\Theta}\right] = \log \sum_{\mathbf{z}\in\mathcal{Z}} \mathbb{P}\left[\mathbf{x}^i, \mathbf{z} \mid \boldsymbol{\Theta}\right]$$

$$= \log \sum_{\mathbf{z}\in\mathcal{Z}} \mathbb{P}\left[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0\right] \cdot \frac{\mathbb{P}\left[\mathbf{x}^i, \mathbf{z} \mid \boldsymbol{\Theta}\right]}{\mathbb{P}\left[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0\right]}$$

$$= \log \mathbb{E}_{\mathbf{z}\sim\mathbb{P}\left[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0\right]} \left[\frac{\mathbb{P}\left[\mathbf{x}^i, \mathbf{z} \mid \boldsymbol{\Theta}\right]}{\mathbb{P}\left[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0\right]}\right]$$

$$\geq \mathbb{E}_{\mathbf{z}\sim\mathbb{P}\left[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0\right]} \log \left[\frac{\mathbb{P}\left[\mathbf{x}^i, \mathbf{z} \mid \boldsymbol{\Theta}\right]}{\mathbb{P}\left[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0\right]}\right]$$

Deriving the EM algorithm

- Assume discrete latent variables (for simplic
- Suppose we already have an estimate $\boldsymbol{\Theta}^0$ with

$$\log \mathbb{P}\left[\mathbf{x}^i \mid \boldsymbol{\Theta}\right] = \log \sum_{\mathbf{z} \in \mathcal{Z}} \mathbb{P}\left[\mathbf{x}^i, \mathbf{z} \mid \boldsymbol{\Theta}\right]$$

$$= \log \sum_{\mathbf{z} \in \mathcal{Z}} \mathbb{P}\left[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0\right] \cdot \frac{\mathbb{P}\left[\mathbf{x}^i, \mathbf{z} \mid \boldsymbol{\Theta}\right]}{\mathbb{P}\left[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0\right]}$$

$$= \log \mathbb{E}_{\mathbf{z} \sim \mathbb{P}\left[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0\right]}\left[\frac{\mathbb{P}\left[\mathbf{x}^i, \mathbf{z} \mid \boldsymbol{\Theta}\right]}{\mathbb{P}\left[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0\right]}\right]$$

$$\geq \mathbb{E}_{\mathbf{z} \sim \mathbb{P}\left[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0\right]} \log \left[\frac{\mathbb{P}\left[\mathbf{x}^i, \mathbf{z} \mid \boldsymbol{\Theta}\right]}{\mathbb{P}\left[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0\right]}\right]$$

Looks like an expectation

$$\mathbb{E}[f(x)] = \sum_{x \in \mathcal{X}} \mathbb{P}[x] \cdot f(x)$$

Just multiplying and dividing by the same quantity

For concave functions

$$\frac{f(x) + f(y)}{2} \leq f\left(\frac{x+y}{2}\right)$$

Jensen's inequality:
For concave functions
$\mathbb{E}[f(x)] \leq f(\mathbb{E}[x])$

Deriving the EM algorithm

- Assume discrete latent variables (for simplicity)
- Suppose we already have an estimate $\boldsymbol{\Theta}^0$ with us somehow

$$\log \mathbb{P}\left[\mathbf{x}^i \mid \boldsymbol{\Theta}\right] \geq \mathbb{E}_{\mathbf{z}\sim\mathbb{P}[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0]} \log\left[\frac{\mathbb{P}\left[\mathbf{x}^i, \mathbf{z} \mid \boldsymbol{\Theta}\right]}{\mathbb{P}[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0]}\right]$$

$$= \mathbb{E}_{\mathbf{z}\sim\mathbb{P}[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0]} \log \mathbb{P}\left[\mathbf{x}^i, \mathbf{z} \mid \boldsymbol{\Theta}\right] - \mathbb{E}_{\mathbf{z}\sim\mathbb{P}[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0]} \log \mathbb{P}\left[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0\right]$$

- Thus, we have

$$\max_{\boldsymbol{\Theta}} \log \mathbb{P}[X \mid \boldsymbol{\Theta}] = \max_{\boldsymbol{\Theta}} \sum_{i=1}^{n} \log \mathbb{P}\left[\mathbf{x}^i \mid \boldsymbol{\Theta}\right]$$

$$\geq \max_{\boldsymbol{\Theta}} \sum_{i=1}^{n} \mathbb{E}_{\mathbf{z}\sim\mathbb{P}[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0]} \log \mathbb{P}\left[\mathbf{x}^i, \mathbf{z} \mid \boldsymbol{\Theta}\right]$$

Deriving the EM algorithm

- Assume discrete latent variables (for simplicity)
- Suppose we already have an estimate $\boldsymbol{\Theta}^0$ with us somehow

$$\log \mathbb{P}\left[\mathbf{x}^i \mid \boldsymbol{\Theta}\right] \geq \mathbb{E}_{\mathbf{z} \sim \mathbb{P}[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0]} \log \left[\frac{\mathbb{P}\left[\mathbf{x}^i, \mathbf{z} \mid \boldsymbol{\Theta}\right]}{\mathbb{P}\left[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0\right]}\right]$$

$$= \mathbb{E}_{\mathbf{z} \sim \mathbb{P}[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0]} \log \mathbb{P}\left[\mathbf{x}^i, \mathbf{z} \mid \boldsymbol{\Theta}\right] - \mathbb{E}_{\mathbf{z} \sim \mathbb{P}[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0]} \log \mathbb{P}\left[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0\right]$$

Does not depend on $\boldsymbol{\Theta}$

- Thus, we have

$$\max_{\boldsymbol{\Theta}} \log \mathbb{P}[X \mid \boldsymbol{\Theta}] = \max_{\boldsymbol{\Theta}} \sum_{i=1}^{n} \log \mathbb{P}\left[\mathbf{x}^i \mid \boldsymbol{\Theta}\right]$$

$$\geq \max_{\boldsymbol{\Theta}} \sum_{i=1}^{n} \mathbb{E}_{\mathbf{z} \sim \mathbb{P}[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0]} \log \mathbb{P}\left[\mathbf{x}^i, \mathbf{z} \mid \boldsymbol{\Theta}\right]$$

Maximizing RHS will improve the data likelihood

Holds true for every $\boldsymbol{\Theta}^0$

If we some $\boldsymbol{\Theta}$ gives a large value for RHS, it will give an even larger value for LHS!

The EM Algorithm

Don't trust a single $\mathbf{z} \in \mathcal{Z}$ but trust them in expectation

EM ALGORITHM

1. Initialize $\boldsymbol{\Theta}^0$
2. For each $i \in [n]$
   1. Let $Q_{i,t}(\boldsymbol{\Theta}) = \mathbb{E}_{\mathbf{z} \sim \mathbb{P}[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^t]} \log \mathbb{P}[\mathbf{x}^i, \mathbf{z} \mid \boldsymbol{\Theta}]$ — E-step
3. Update $\boldsymbol{\Theta}^{t+1} = \arg\max_{\boldsymbol{\Theta}} \sum_{i=1}^n Q_{i,t}(\boldsymbol{\Theta})$ — M-step
4. Repeat until convergence

- Notice that if we let $\mathbf{z}^{i,t} = \arg\max_{\mathbf{z} \in \mathcal{Z}} \mathbb{P}\left[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^t\right]$
- … and use $Q_{i,t}(\boldsymbol{\Theta}) = \log \mathbb{P}[\mathbf{x}^i, \mathbf{z}^{i,t} \mid \boldsymbol{\Theta}]$ (no expectations)
- … then we get hard AltOpt ☺

The EM algorithm implements soft AltOpt

- For sake of simplicity, let $\mathcal{Z} = \{1, 2, \ldots, K\}$ (GMM, MR)
- Below we reproduce EM exactly for the above special case

EM ALGORITHM

1. Initialize $\boldsymbol{\Theta}^0$
2. For each $i \in [n]$
    1. For each $k \in [K]$
        1. Let $\gamma^{i,k,t} = \mathbb{P}\left[\mathbf{z}^k \mid \mathbf{x}^i, \boldsymbol{\Theta}^{\mathrm{t}}\right]$
3. Update $\boldsymbol{\Theta}^{t+1} = \arg\max_{\boldsymbol{\Theta}} \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma^{i,k,t} \cdot \log \mathbb{P}\left[\mathbf{x}^i, \mathbf{z}^k \mid \boldsymbol{\Theta}\right]$
4. Repeat until convergence

Weighted MLE!! Exactly what we did for GMM, MR

Exercise: verify that EM exactly recovers soft AltOpt for GMM, MR, PPCA

Why the EM algorithm is the way it is

- Basically, the message EM gives us is

"Working with $\log \mathbb{P}\left[\mathbf{x}^i \mid \boldsymbol{\Theta}\right]$ *makes life difficult,*
Working with $\mathbb{E}_{\mathbf{z}\sim\mathbb{P}[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0]} \log\left[\frac{\mathbb{P}[\mathbf{x}^i, \mathbf{z} \mid \boldsymbol{\Theta}]}{\mathbb{P}[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0]}\right]$ *gives peace"*

- One reason why EM suggests this is because

$$\mathbb{E}_{\mathbf{z}\sim\mathbb{P}[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0]} \log\left[\frac{\mathbb{P}\left[\mathbf{x}^i, \mathbf{z} \mid \boldsymbol{\Theta}^0\right]}{\mathbb{P}\left[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0\right]}\right]$$

$$= \mathbb{E}_{\mathbf{z}\sim\mathbb{P}[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0]} \log\left[\frac{\mathbb{P}\left[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0\right] \cdot \mathbb{P}\left[\mathbf{x}^i \mid \boldsymbol{\Theta}^0\right]}{\mathbb{P}\left[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0\right]}\right]$$

$$= \mathbb{E}_{\mathbf{z}\sim\mathbb{P}[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0]} \log\left[\mathbb{P}\left[\mathbf{x}^i \mid \boldsymbol{\Theta}^0\right]\right]$$

$$= \log \mathbb{P}\left[\mathbf{x}^i \mid \boldsymbol{\Theta}^0\right]$$

Why the EM algorithm is the way it is

- Basically, the message EM gives us is

"Working with $\log \mathbb{P}\left[\mathbf{x}^i \mid \boldsymbol{\Theta}\right]$ *makes life difficult,*
Working with $\mathbb{E}_{\mathbf{z} \sim \mathbb{P}[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0]} \log \left[\frac{\mathbb{P}[\mathbf{x}^i, \mathbf{z} \mid \boldsymbol{\Theta}]}{\mathbb{P}[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0]}\right]$ *gives peace"*

- Nice! Let us denote

$$Q_t(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \mathbb{E}_{\mathbf{z} \sim \mathbb{P}[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^t]} \log \left[\frac{\mathbb{P}[\mathbf{x}^i, \mathbf{z} \mid \boldsymbol{\Theta}]}{\mathbb{P}[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^t]}\right]$$

- This means $Q_t(\boldsymbol{\Theta}^t) = \log \mathbb{P}[X \mid \boldsymbol{\Theta}^t]$
- But since the M-step maximizes the function $Q_t(\cdot)$, we must have

$$Q_t(\boldsymbol{\Theta}^{t+1}) \geq Q_t(\boldsymbol{\Theta}^t) = \log \mathbb{P}[X \mid \boldsymbol{\Theta}^t]$$

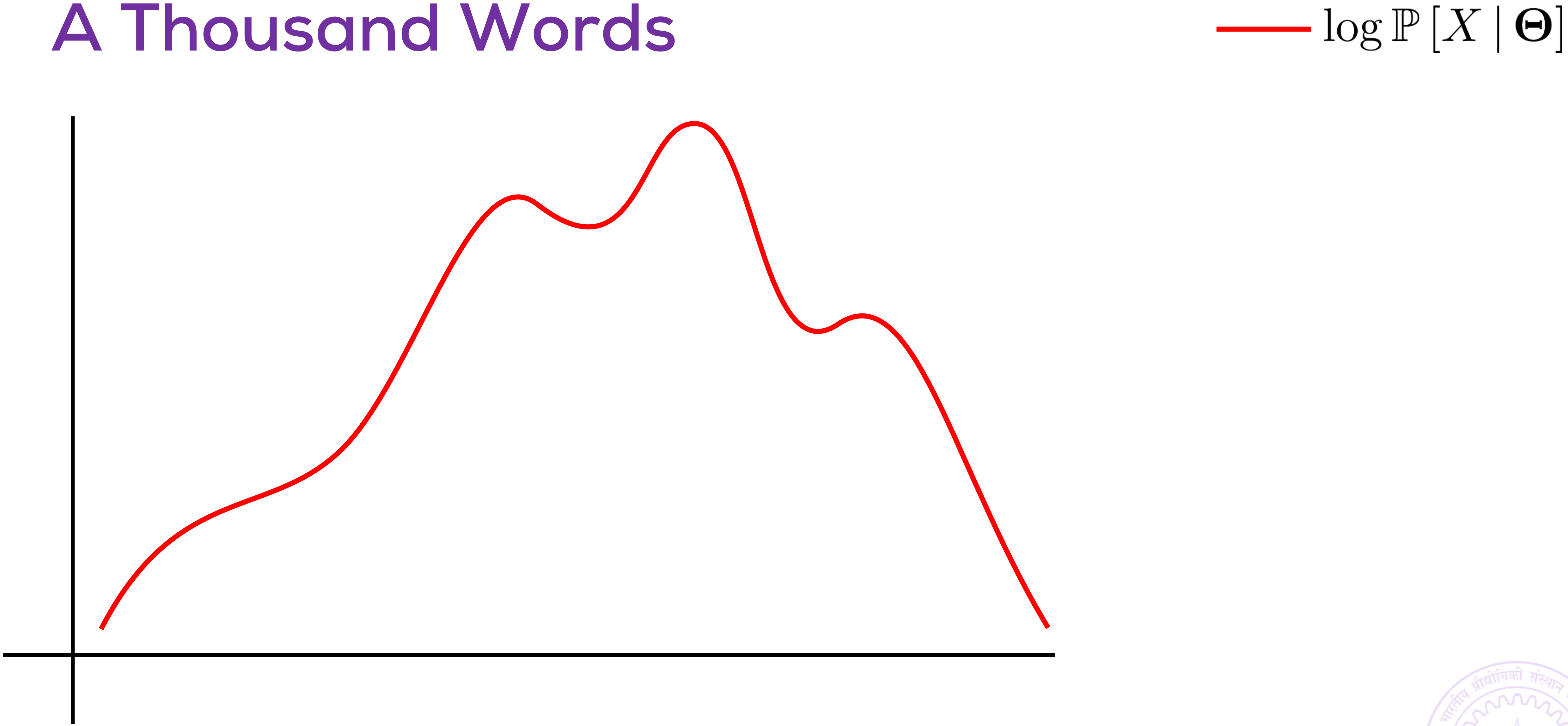
Why the EM algorithm is the way it is

- So we have
$$Q_t(\boldsymbol{\Theta}^{t+1}) \geq Q_t(\boldsymbol{\Theta}^t) = \log \mathbb{P}[X \mid \boldsymbol{\Theta}^t]$$
- But remember that we showed that for any $\boldsymbol{\Theta}$ and any $\boldsymbol{\Theta}^0$
$$\log \mathbb{P}\left[\mathbf{x}^i \mid \boldsymbol{\Theta}\right] \geq \mathbb{E}_{\mathbf{z}\sim\mathbb{P}[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0]} \log \left[\frac{\mathbb{P}\left[\mathbf{x}^i, \mathbf{z} \mid \boldsymbol{\Theta}\right]}{\mathbb{P}[\mathbf{z} \mid \mathbf{x}^i, \boldsymbol{\Theta}^0]}\right]$$
- This means for every $\boldsymbol{\Theta}$, we have $\log \mathbb{P}[X \mid \boldsymbol{\Theta}] \geq Q_t(\boldsymbol{\Theta})$
- Very very nice since this means
$$\log \mathbb{P}[X \mid \boldsymbol{\Theta}^{t+1}] \geq Q_t(\boldsymbol{\Theta}^{t+1}) \geq Q_t(\boldsymbol{\Theta}^t) = \log \mathbb{P}[X \mid \boldsymbol{\Theta}^t]$$
- The EM algorithm never decreases the log likelihood!
- Hard AltOpt does not have such guarantees in general
- Monotonic progress but may converge to local optimum ☹

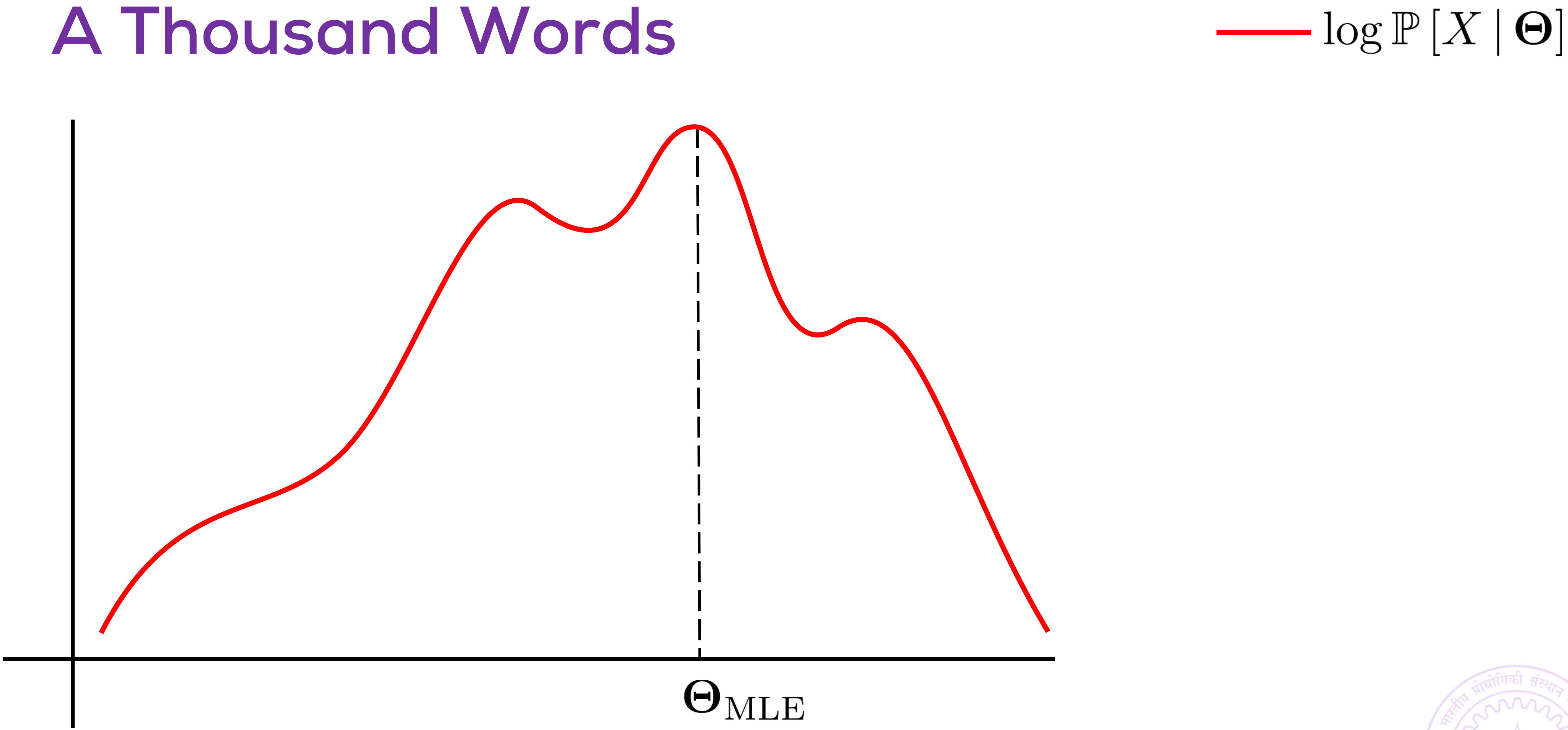
A Thousand Words

A Thousand Words

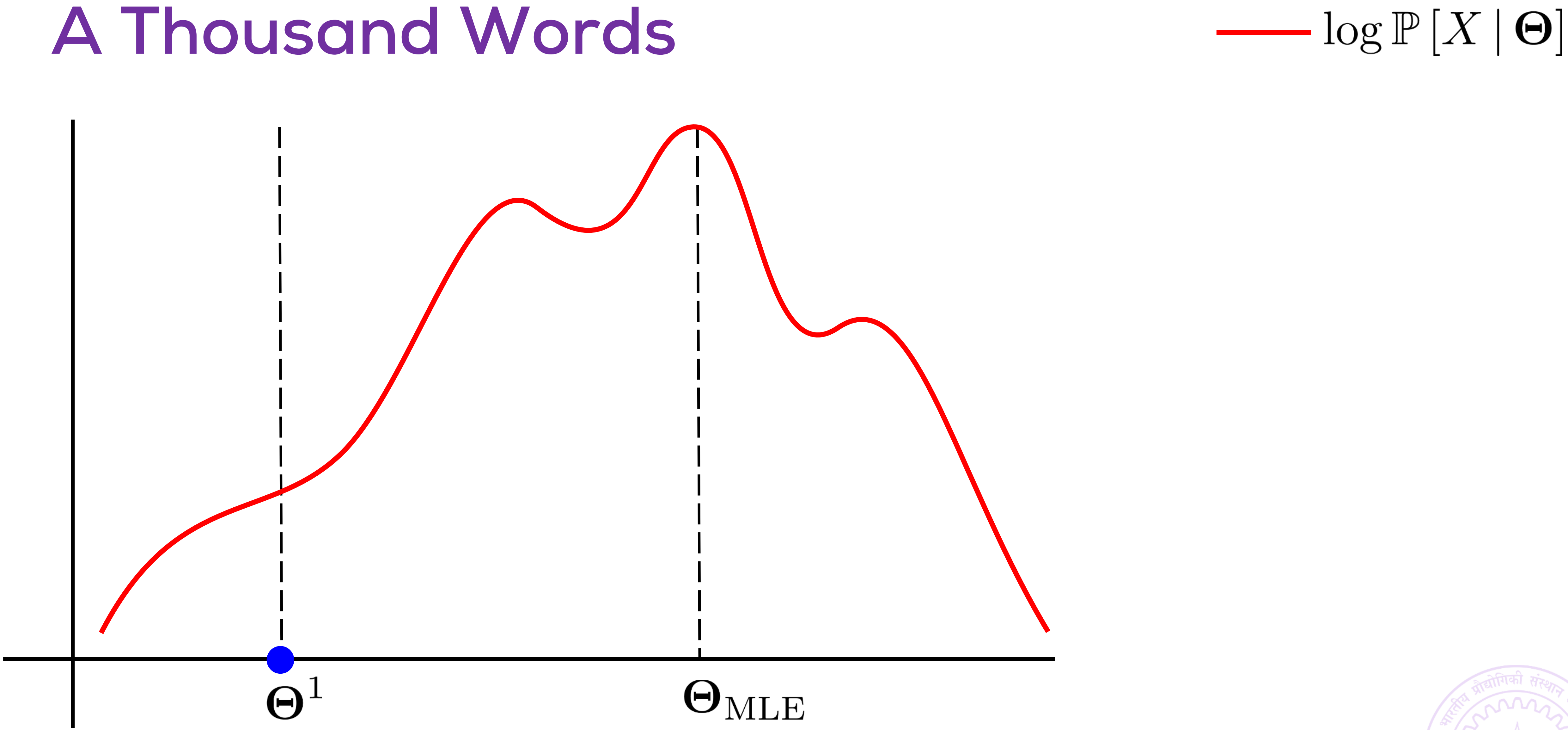
A Thousand Words



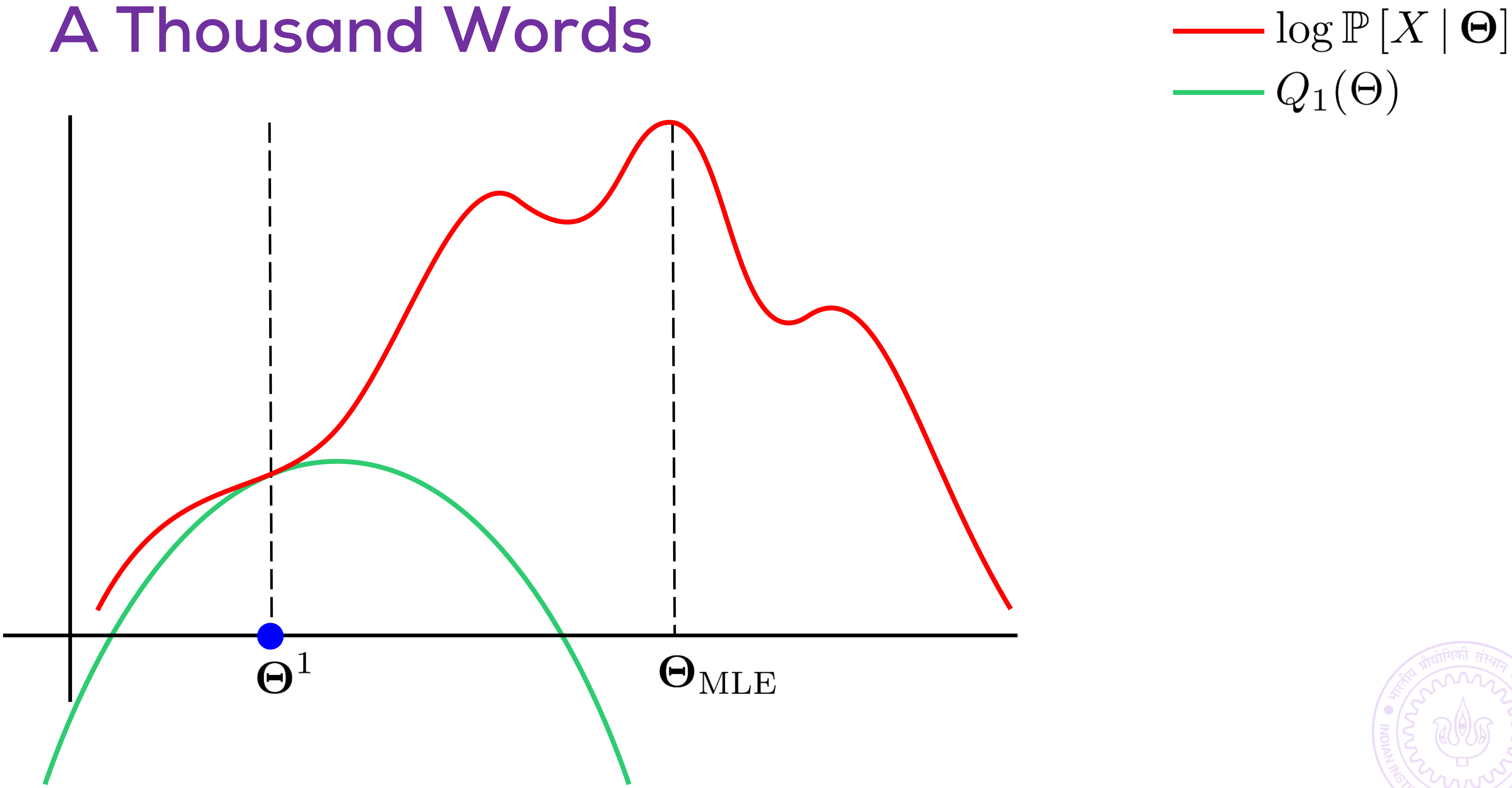
A Thousand Words



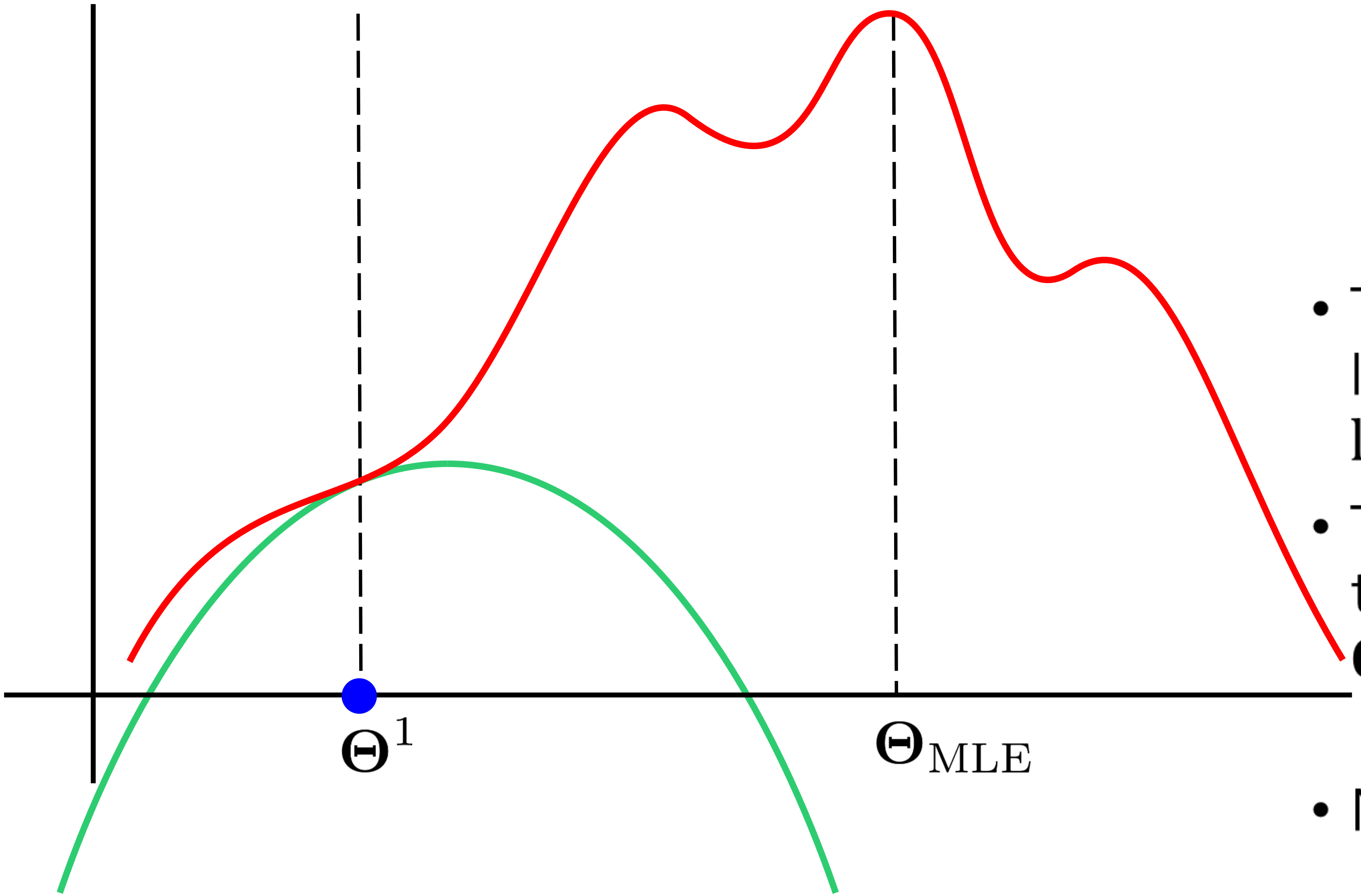
A Thousand Words



A Thousand Words



A Thousand Words



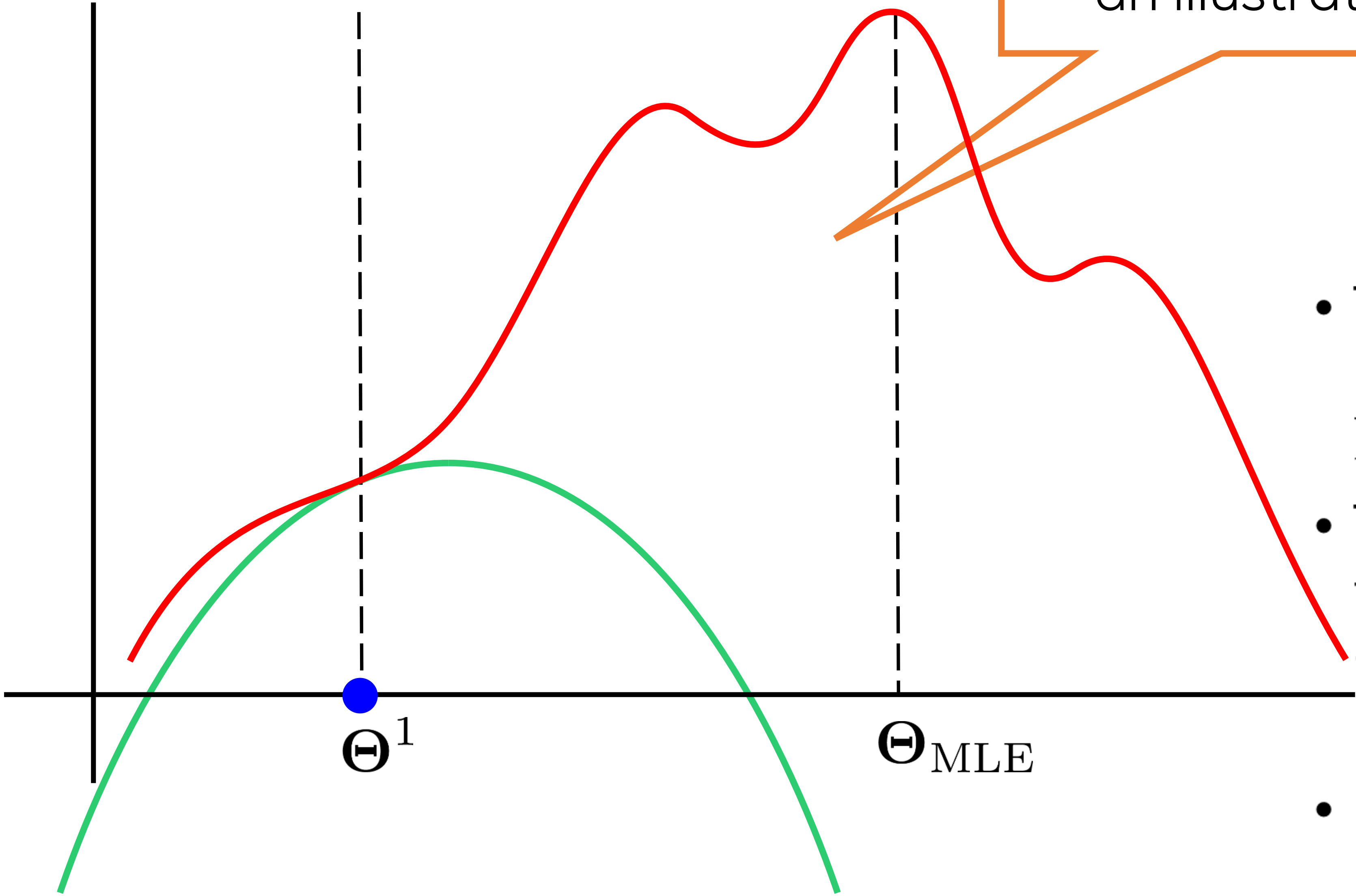
- The Q_{t}-curves always lie below the red curve $\log \mathbb{P}[X \mid \boldsymbol{\Theta}] \geq Q_t(\boldsymbol{\Theta}), \forall \boldsymbol{\Theta}$
- The Q_t curves always touch the red curve at $\boldsymbol{\Theta}^t$ because $Q_t(\boldsymbol{\Theta}^t) = \log \mathbb{P}[X \mid \boldsymbol{\Theta}^t]$
- M-step maximizes $Q_t(\cdot)$

A Thousand Words

$Q_t(\cdot)$ is not always a quadratic fn. Just an illustration

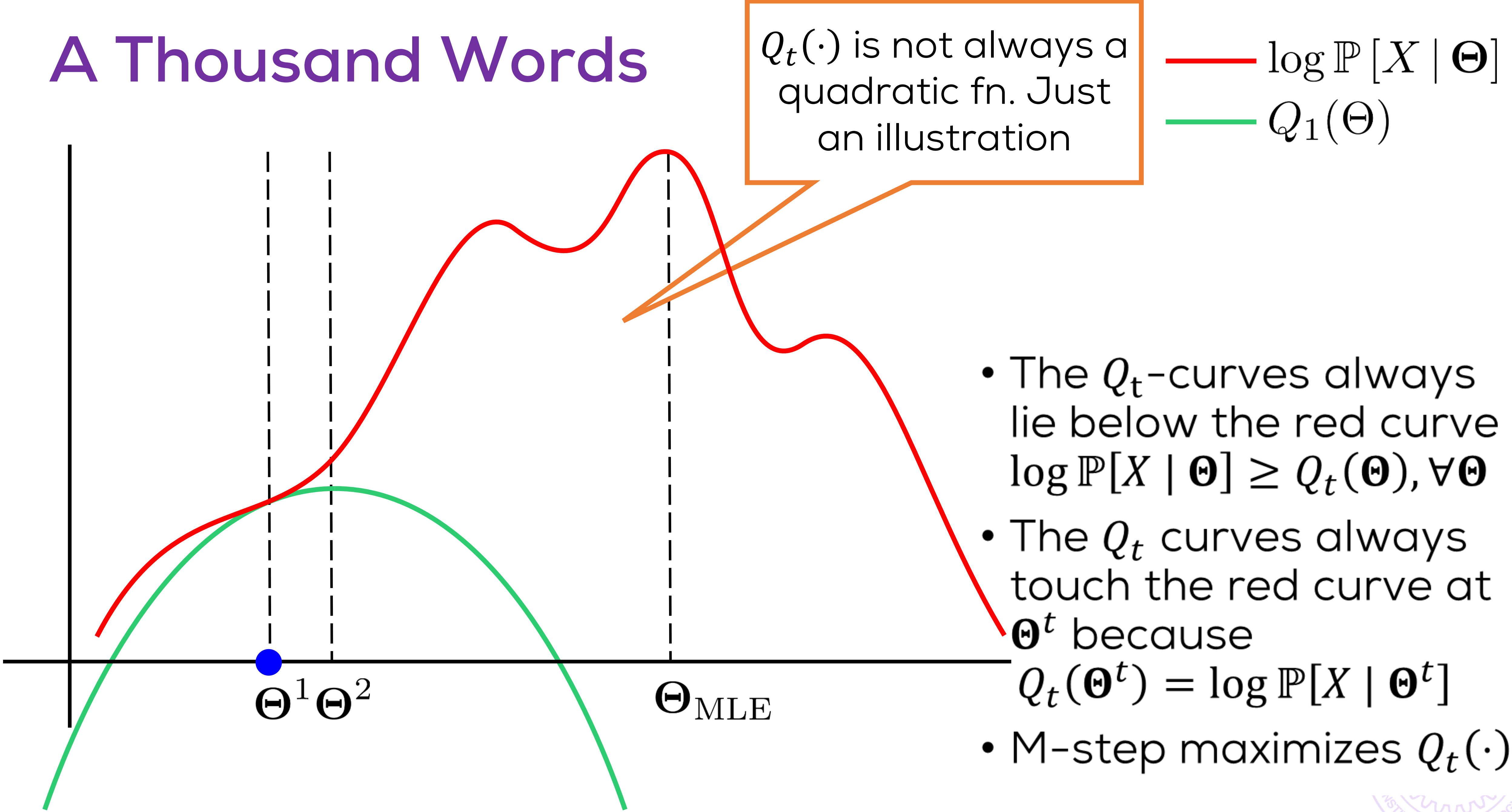
$\log \mathbb{P}\left[X \mid \mathbf{\Theta}\right]$

$Q_1(\Theta)$



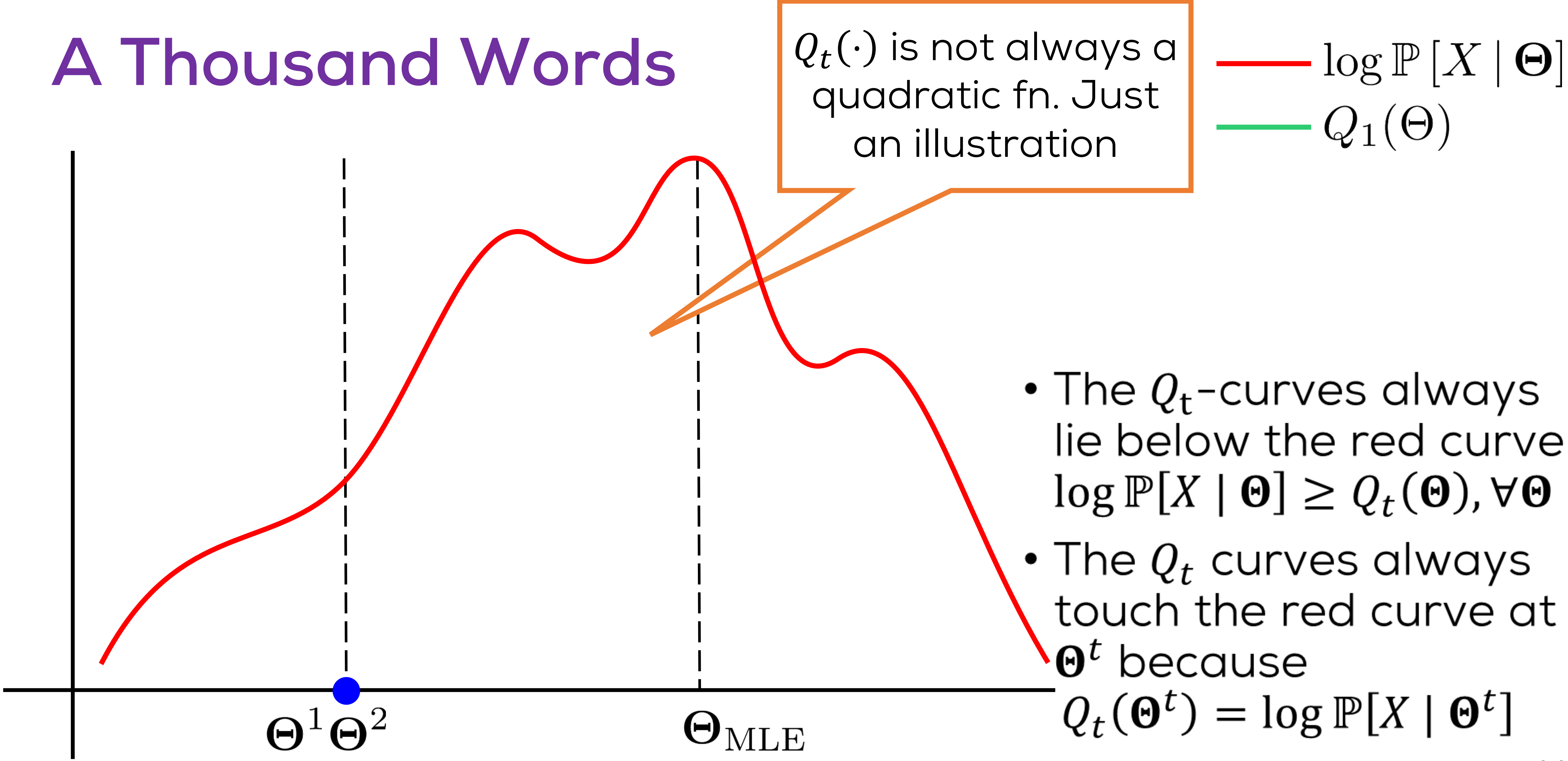
- The Q_{t}-curves always lie below the red curve $\log \mathbb{P}[X \mid \mathbf{\Theta}] \geq Q_t(\mathbf{\Theta}), \forall \mathbf{\Theta}$
- The Q_t curves always touch the red curve at $\mathbf{\Theta}^t$ because $Q_t(\mathbf{\Theta}^t) = \log \mathbb{P}[X \mid \mathbf{\Theta}^t]$
- M-step maximizes $Q_t(\cdot)$

A Thousand Words



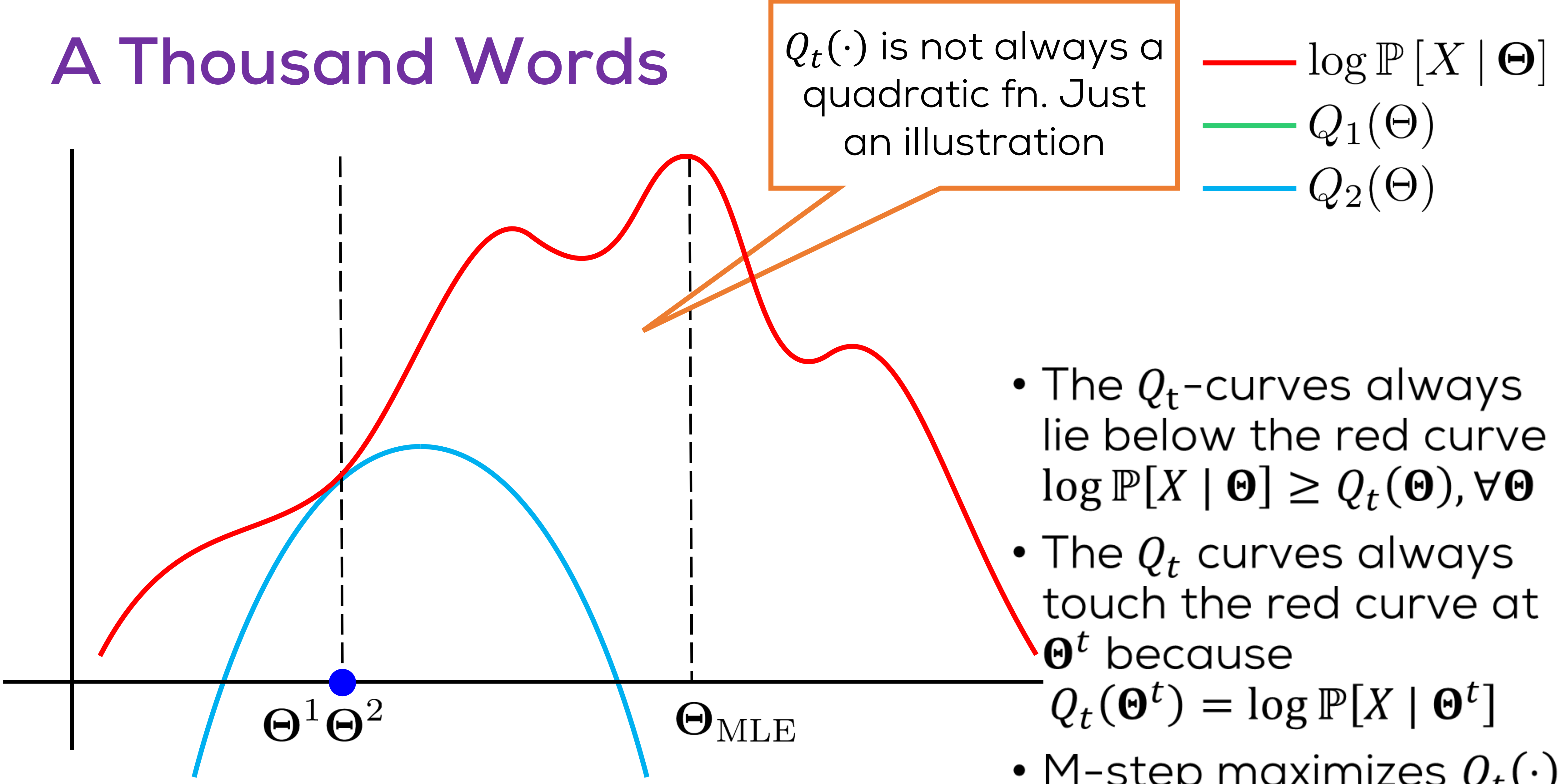
- The Q_t-curves always lie below the red curve $\log \mathbb{P}[X \mid \mathbf{\Theta}] \geq Q_t(\mathbf{\Theta}), \forall \mathbf{\Theta}$
- The Q_t curves always touch the red curve at $\mathbf{\Theta}^t$ because $Q_t(\mathbf{\Theta}^t) = \log \mathbb{P}[X \mid \mathbf{\Theta}^t]$
- M-step maximizes $Q_t(\cdot)$

A Thousand Words



- The Q_t-curves always lie below the red curve $\log \mathbb{P}[X \mid \mathbf{\Theta}] \geq Q_t(\mathbf{\Theta}), \forall \mathbf{\Theta}$
- The Q_t curves always touch the red curve at $\mathbf{\Theta}^t$ because $Q_t(\mathbf{\Theta}^t) = \log \mathbb{P}[X \mid \mathbf{\Theta}^t]$
- M-step maximizes $Q_t(\cdot)$

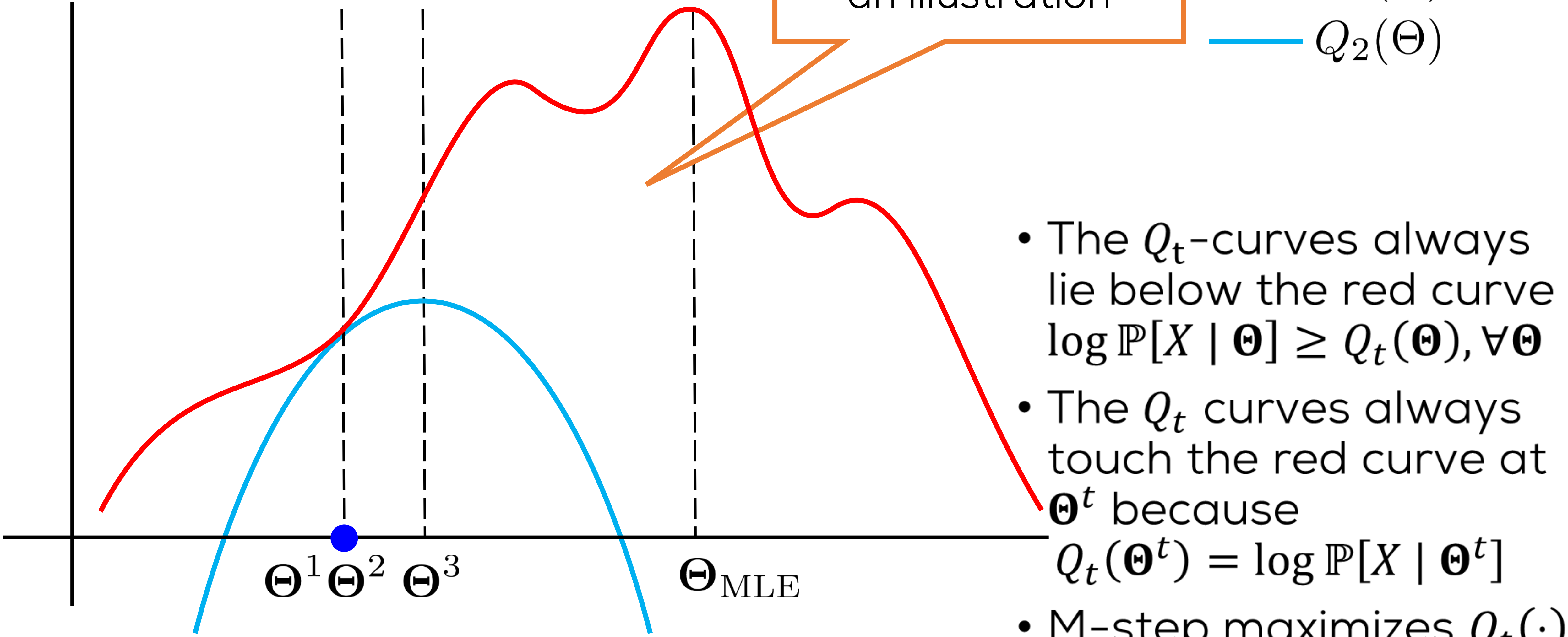
A Thousand Words



- The Q_t-curves always lie below the red curve $\log \mathbb{P}[X \mid \mathbf{\Theta}] \geq Q_t(\mathbf{\Theta}), \forall \mathbf{\Theta}$
- The Q_t curves always touch the red curve at $\mathbf{\Theta}^t$ because $Q_t(\mathbf{\Theta}^t) = \log \mathbb{P}[X \mid \mathbf{\Theta}^t]$
- M-step maximizes $Q_t(\cdot)$

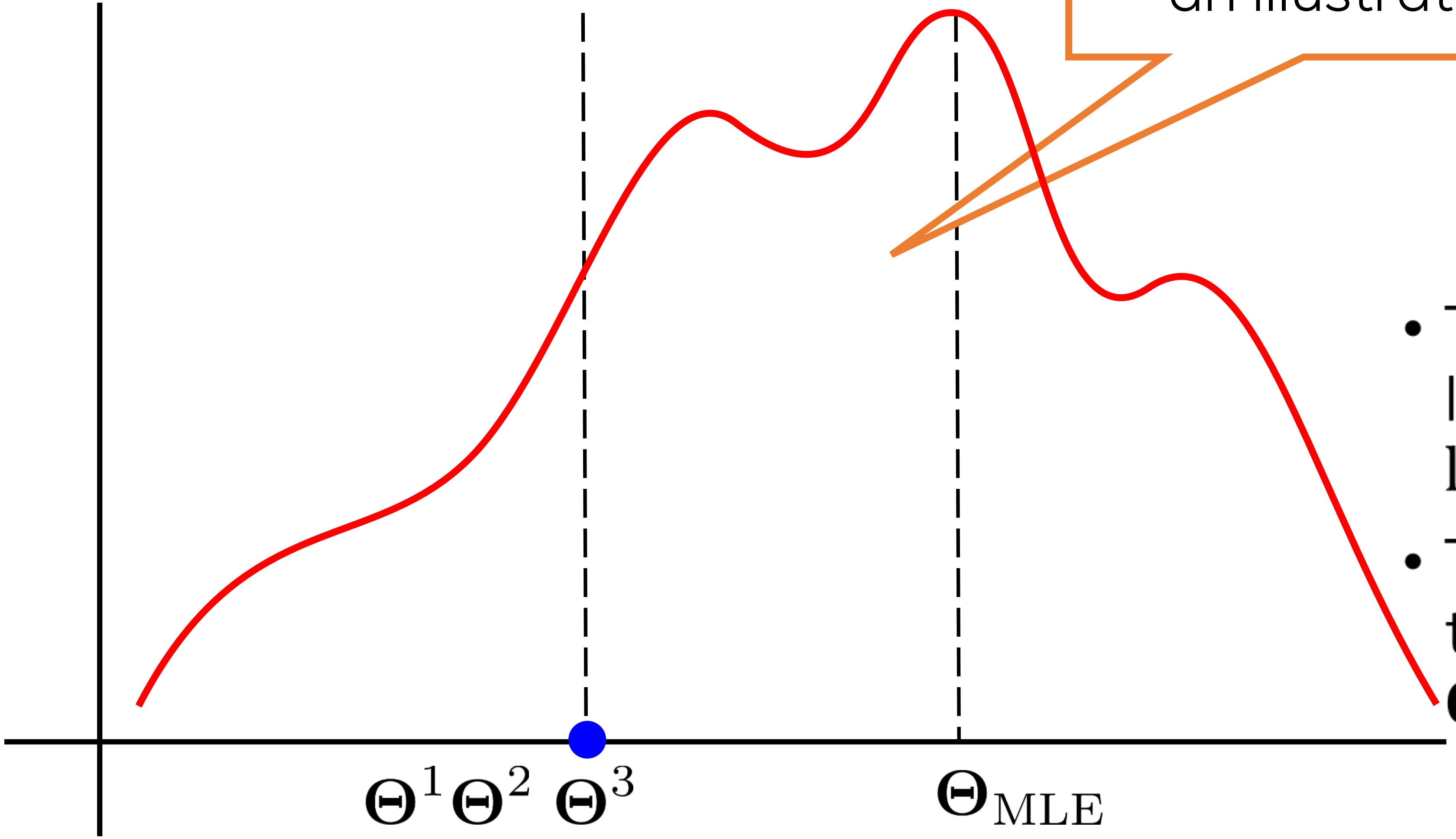
A Thousand Words

$Q_t(\cdot)$ is not always a quadratic fn. Just an illustration



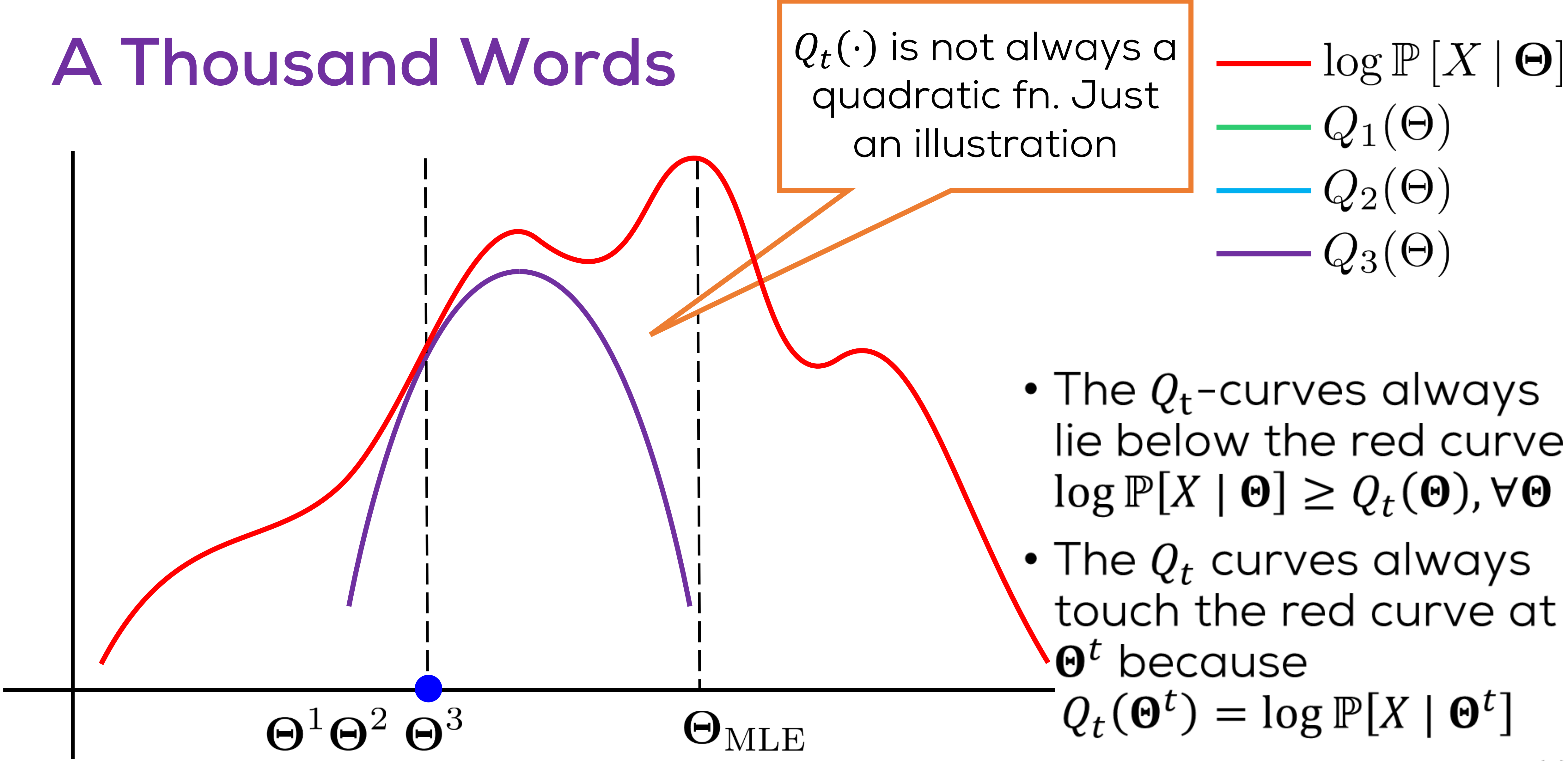
- The Q_t-curves always lie below the red curve $\log \mathbb{P}[X \mid \boldsymbol{\Theta}] \geq Q_t(\boldsymbol{\Theta}), \forall \boldsymbol{\Theta}$
- The Q_t curves always touch the red curve at $\boldsymbol{\Theta}^t$ because $Q_t(\boldsymbol{\Theta}^t) = \log \mathbb{P}[X \mid \boldsymbol{\Theta}^t]$
- M-step maximizes $Q_t(\cdot)$

A Thousand Words



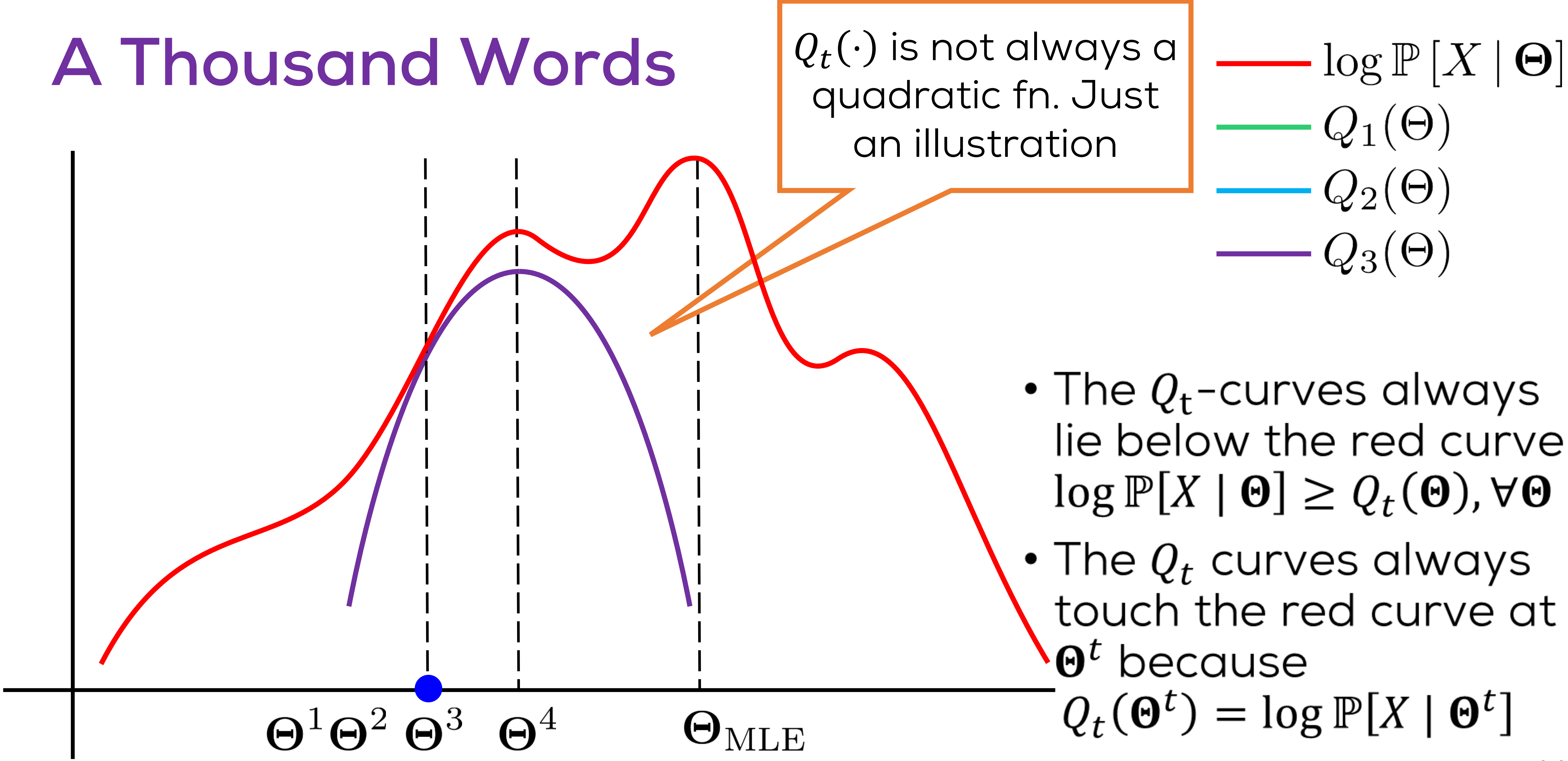
- The Q_t-curves always lie below the red curve $\log \mathbb{P}[X \mid \Theta] \geq Q_t(\Theta), \forall \Theta$
- The Q_t curves always touch the red curve at Θ^t because $Q_t(\Theta^t) = \log \mathbb{P}[X \mid \Theta^t]$
- M-step maximizes $Q_t(\cdot)$

A Thousand Words



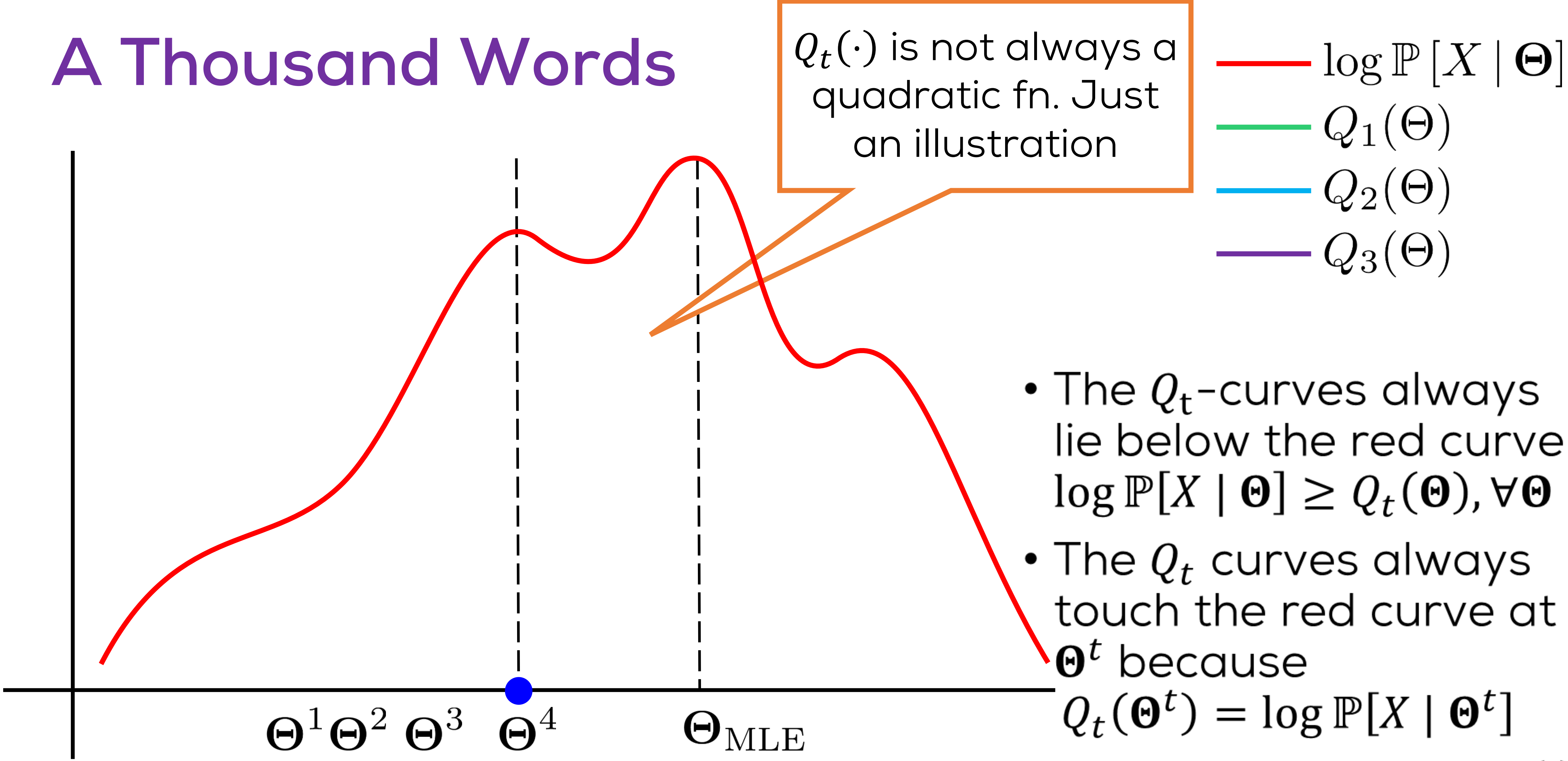
- The Q_t-curves always lie below the red curve $\log \mathbb{P}[X \mid \boldsymbol{\Theta}] \geq Q_t(\boldsymbol{\Theta}), \forall \boldsymbol{\Theta}$
- The Q_t curves always touch the red curve at $\boldsymbol{\Theta}^t$ because $Q_t(\boldsymbol{\Theta}^t) = \log \mathbb{P}[X \mid \boldsymbol{\Theta}^t]$
- M-step maximizes $Q_t(\cdot)$

A Thousand Words



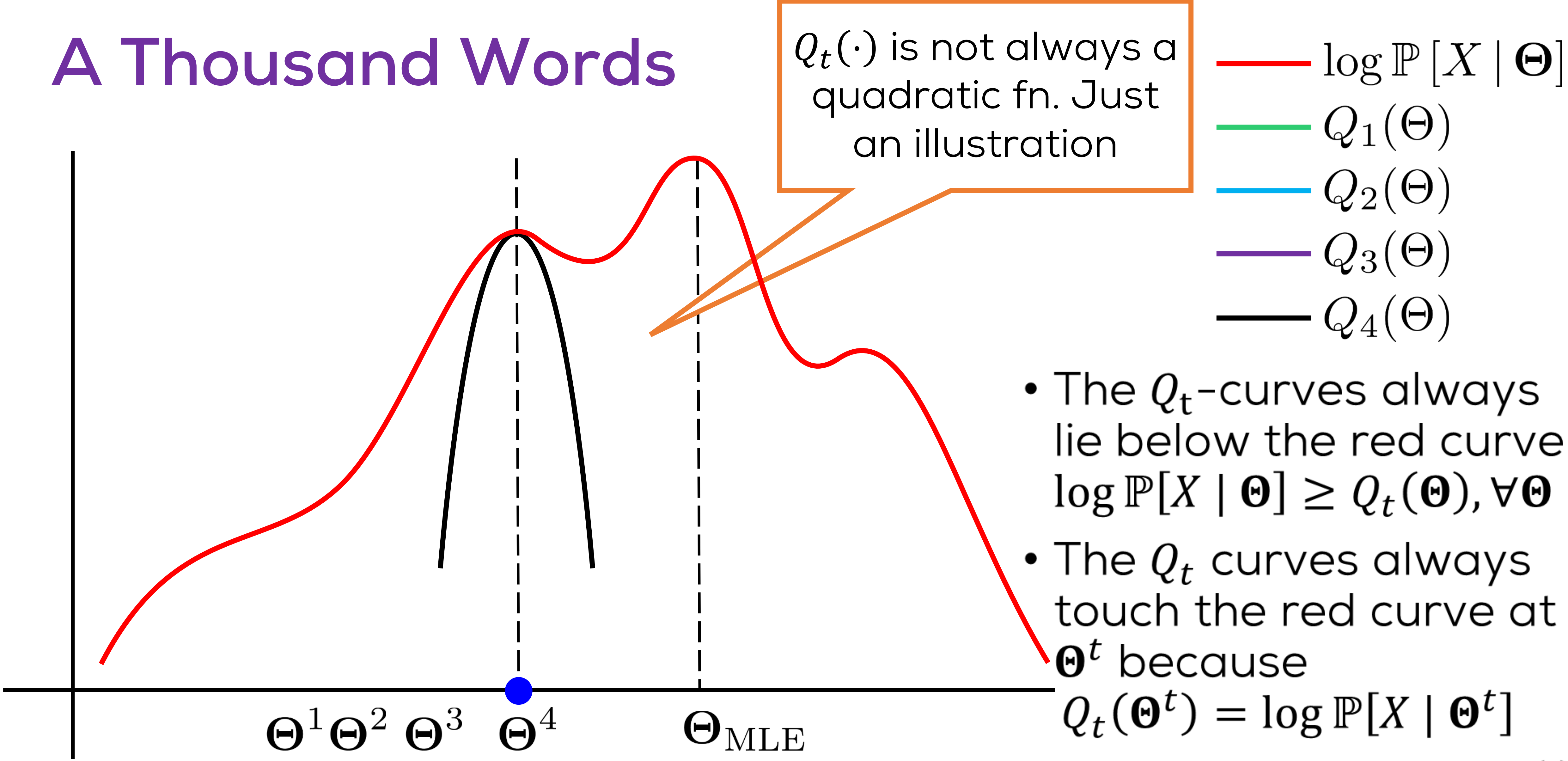
- The Q_t-curves always lie below the red curve $\log \mathbb{P}[X \mid \mathbf{\Theta}] \geq Q_t(\mathbf{\Theta}), \forall \mathbf{\Theta}$
- The Q_t curves always touch the red curve at $\mathbf{\Theta}^t$ because $Q_t(\mathbf{\Theta}^t) = \log \mathbb{P}[X \mid \mathbf{\Theta}^t]$
- M-step maximizes $Q_t(\cdot)$

A Thousand Words



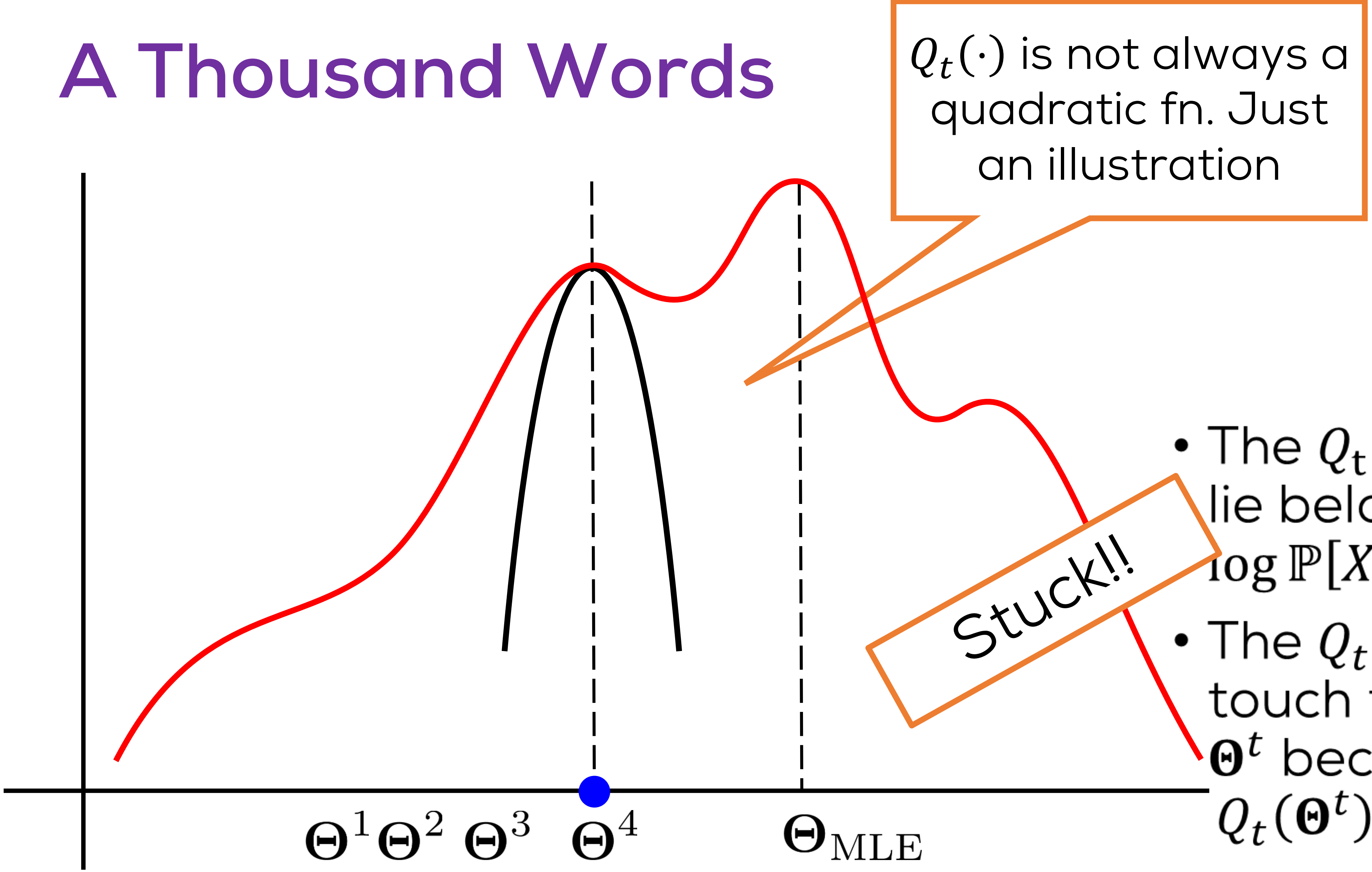
- The Q_t-curves always lie below the red curve $\log \mathbb{P}[X \mid \boldsymbol{\Theta}] \geq Q_t(\boldsymbol{\Theta}), \forall \boldsymbol{\Theta}$
- The Q_t curves always touch the red curve at $\boldsymbol{\Theta}^t$ because $Q_t(\boldsymbol{\Theta}^t) = \log \mathbb{P}[X \mid \boldsymbol{\Theta}^t]$
- M-step maximizes $Q_t(\cdot)$

A Thousand Words



- The Q_t-curves always lie below the red curve $\log \mathbb{P}[X \mid \mathbf{\Theta}] \geq Q_t(\mathbf{\Theta}), \forall \mathbf{\Theta}$
- The Q_t curves always touch the red curve at $\mathbf{\Theta}^t$ because $Q_t(\mathbf{\Theta}^t) = \log \mathbb{P}[X \mid \mathbf{\Theta}^t]$
- M-step maximizes $Q_t(\cdot)$

A Thousand Words



- The Q_t-curves always lie below the red curve $\log \mathbb{P}[X \mid \boldsymbol{\Theta}] \geq Q_t(\boldsymbol{\Theta}), \forall \boldsymbol{\Theta}$
- The Q_t curves always touch the red curve at $\boldsymbol{\Theta}^t$ because $Q_t(\boldsymbol{\Theta}^t) = \log \mathbb{P}[X \mid \boldsymbol{\Theta}^t]$
- M-step maximizes $Q_t(\cdot)$

Some Thoughts

- EM is useful whenever we have latent variables
- Missing data can be thought of as latent variables
- May variants of EM exist
  - “Fully corrective” E and M steps
  - Incomplete/partial E and M steps
  - Stochastic E and M steps
- More advanced extensions exist (e.g. variational Bayes)
- See the paper by Balakrishnan, Wainwright, Yu
  Statistical guarantees for the EM algorithm: From population to sample-based analysis, Annals of Statistics 45(1): 77-120, 2017.

Please give your Feedback

http://tinyurl.com/ml17-18afb