Expectation-Maximization

EECE454 Intro. to Machine Learning Systems

Fall 2024

Recap

- **GMM.** We fit a Gaussian mixture density function to the training data

$$p(\mathbf{x} \mid \theta) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(\mathbf{x} \mid \mu_k, \Sigma_k)$$

Recap

- **GMM.** We fit a Gaussian mixture density function to the training data

$$p(\mathbf{x} \mid \theta) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(\mathbf{x} \mid \mu_k, \Sigma_k)$$

- The optimization can be done by alternating two steps
  - Special version of EM

1. Initialize $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$.
2. *E-step:* Evaluate responsibilities r_{nk} for every data point $\boldsymbol{x}_n$ using current parameters $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$:

$$r_{nk} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \tag{11.53}$$

3. *M-step:* Reestimate parameters $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ using the current responsibilities r_{nk} (from E-step):

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \boldsymbol{x}_n, \tag{11.54}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top, \tag{11.55}$$

$$\pi_k = \frac{N_k}{N}. \tag{11.56}$$

Recap

- **GMM.** We fit a Gaussian mixture density function to the training data

$$p(\mathbf{x} \mid \theta) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(\mathbf{x} \mid \mu_k, \Sigma_k)$$

  - The optimization can be done by alternating two steps
    - Special version of EM

- **Today.** We take a look at EM in a more general sense
  - Description
  - Convergence

1. Initialize $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$.
2. *E-step:* Evaluate responsibilities r_{nk} for every data point $\boldsymbol{x}_n$ using current parameters $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$:

$$r_{nk} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} . \tag{11.53}$$

3. *M-step:* Reestimate parameters $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ using the current responsibilities r_{nk} (from E-step):

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \boldsymbol{x}_n , \tag{11.54}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top , \tag{11.55}$$

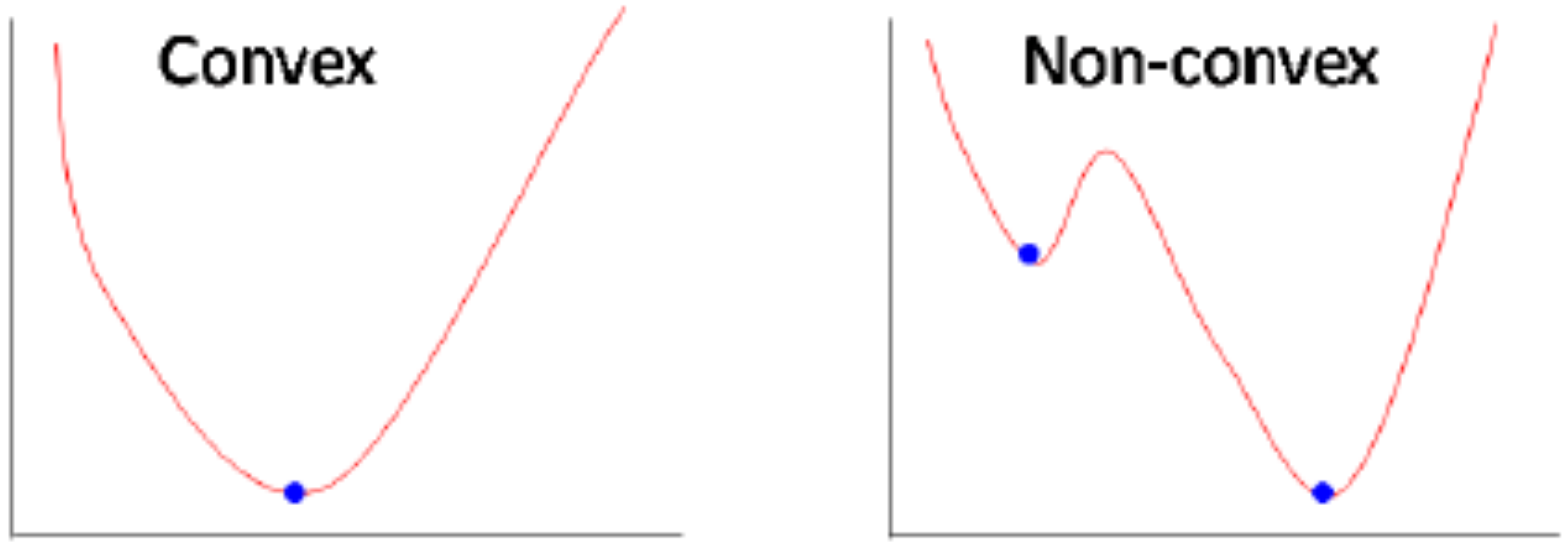
$$\pi_k = \frac{N_k}{N} . \tag{11.56}$$

Prerequisite: Convexity

Convex function

- Before we begin, we briefly familiarize ourselves with the notion of convexity.

Definition (narrow). A differentiable function $f(x) : \mathbb{R} \to \mathbb{R}$ is convex whenever $f''(x) \geq 0$



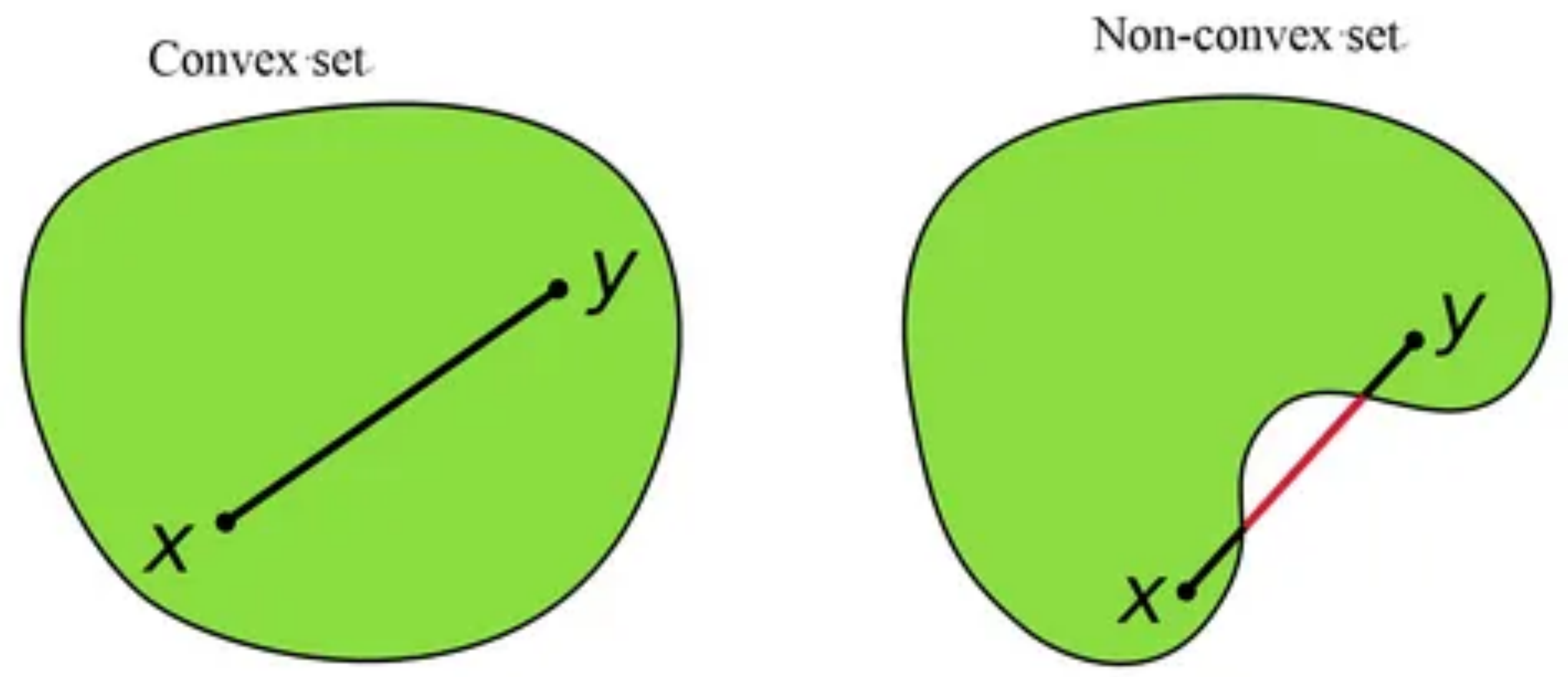
Convex function

- Before we begin, we briefly familiarize ourselves with the notion of convexity.

Definition (narrow). A differentiable function $f(x) : \mathbb{R} \to \mathbb{R}$ is convex whenever $f''(x) \geq 0$

Definition (general). A set $\mathcal{S}$ is a convex set whenever for any $x, y \in \mathcal{S}$, we have

$$(1-\lambda)x + \lambda y \in \mathcal{S}, \qquad \forall \lambda \in [0,1]$$



Convex function

- Before we begin, we briefly familiarize ourselves with the notion of convexity.

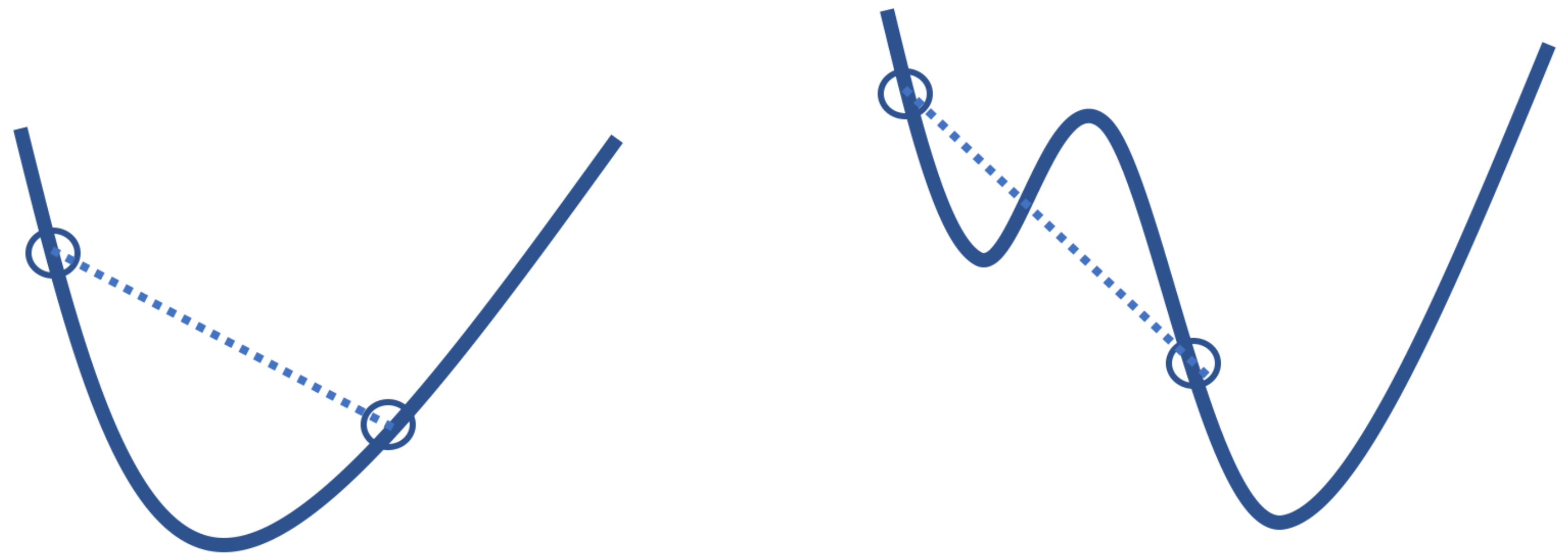
Definition (narrow). A differentiable function $f(x) : \mathbb{R} \to \mathbb{R}$ is convex whenever $f''(x) \geq 0$

Definition (general). A set $\mathcal{S}$ is a convex set whenever for any $x, y \in \mathcal{S}$, we have

$$(1-\lambda)x + \lambda y \in \mathcal{S}, \qquad \forall \lambda \in [0,1]$$

A function $f : \mathcal{S} \to \mathbb{R}$ is a convex function whenever

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$$



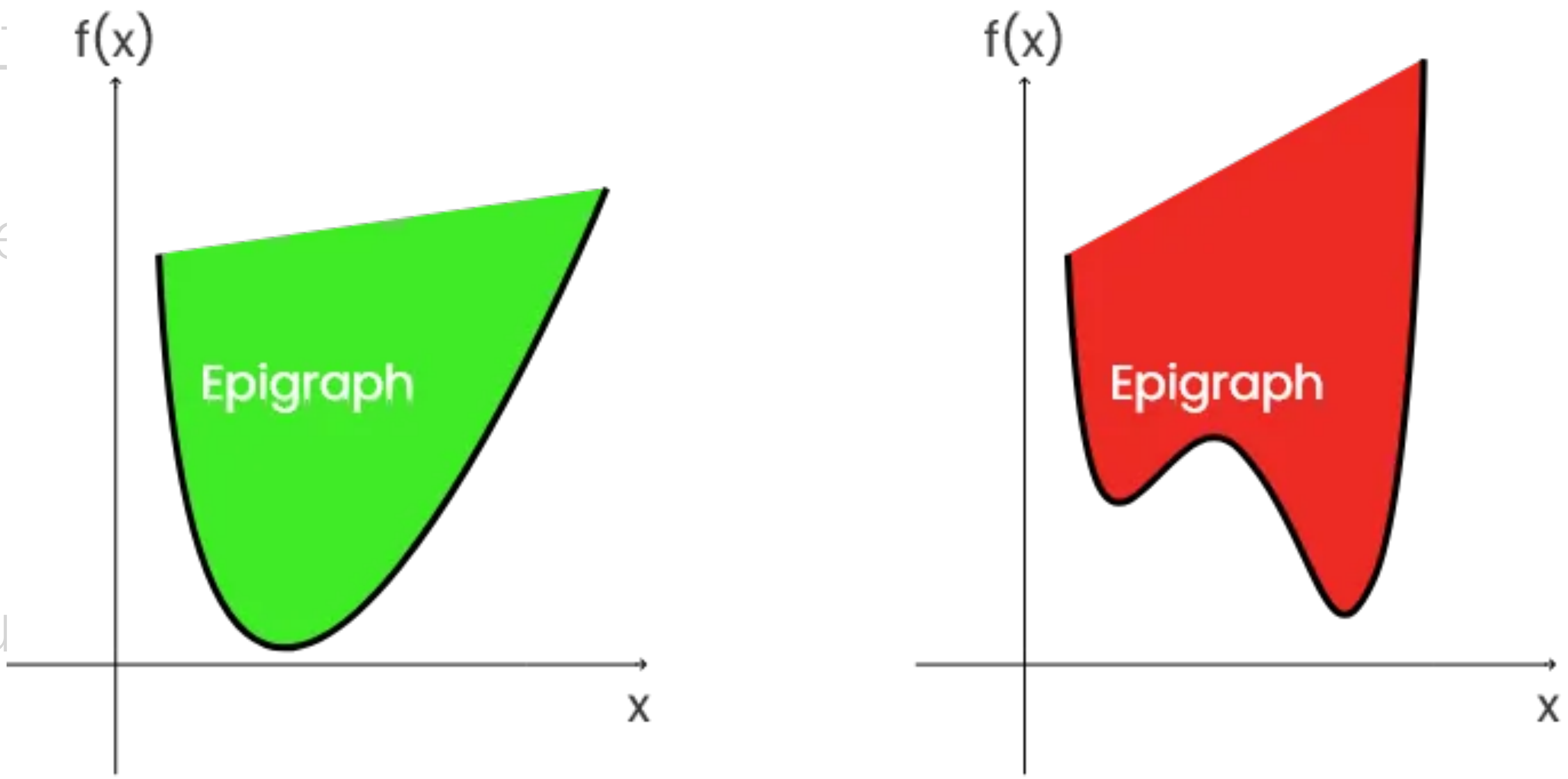
Convex function

- Before we begin, we briefly familiarize ourselves with the notion of convexity.

Definition (narrow). A di ver $f''(x) \geq 0$

Definition (general). A se have

A fu



A convex function Not a convex function

A function is convex if and only if its epigraph is a convex set

Jensen's inequality

- For convex functions, we have a convenient property called Jensen's inequality.

Theorem. Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function, and let X be a random variable. Then, we have

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

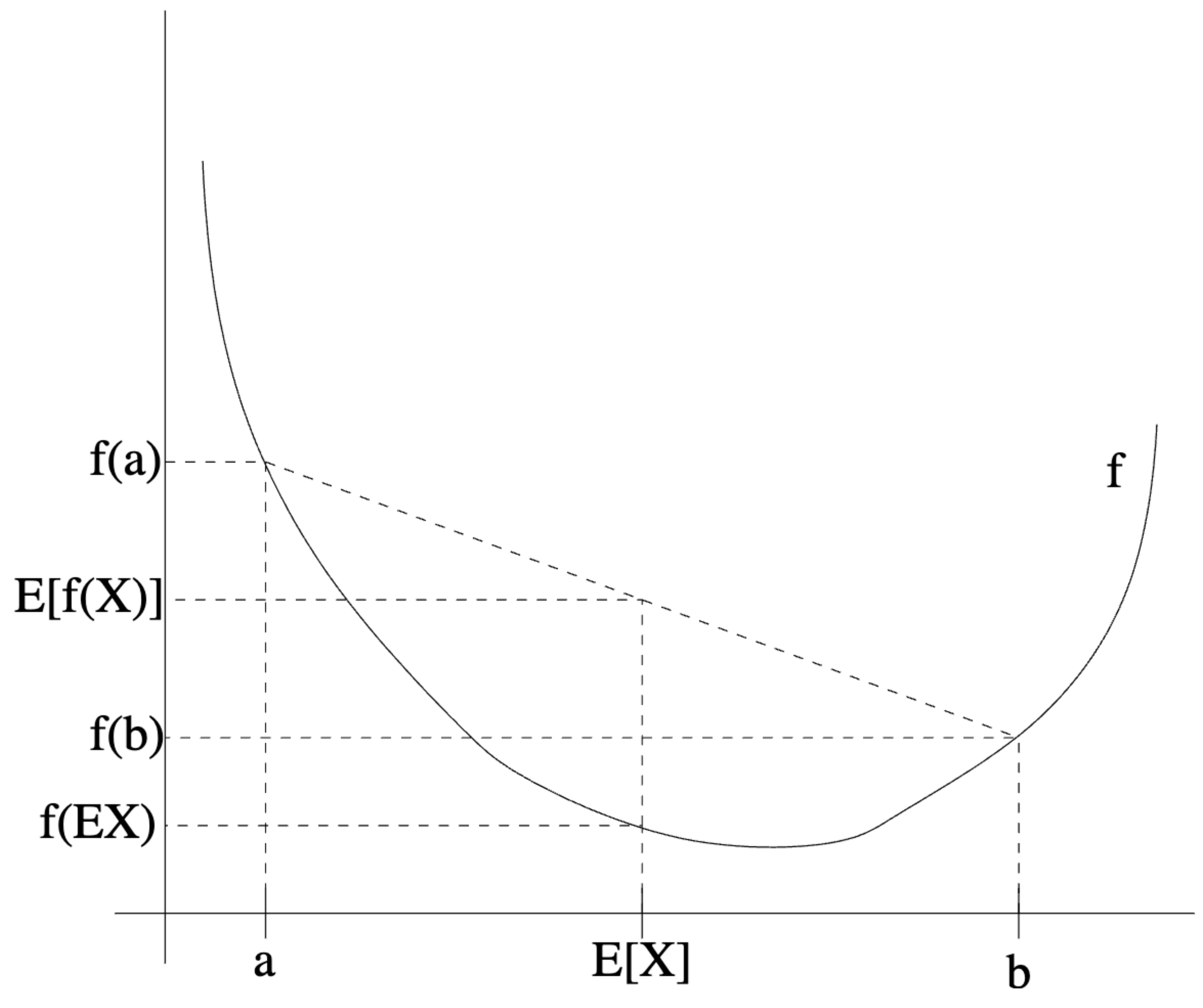
Jensen's inequality

- For convex functions, we have a convenient property called Jensen's inequality.

Theorem. Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function, and let X be a random variable. Then, we have

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

- Proof idea. Expectation = weighted sum = linear combination



Jensen's inequality

- For convex functions, we have a convenient property called Jensen's inequality.

Theorem. Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function, and let X be a random variable. Then, we have

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

- Proof idea. Expectation = weighted sum = linear combination

- Remark. The inequality $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$ holds with equality if $X = \mathbb{E}[X]$ with probability 1.

Expectation-Maximization

Setup

- Suppose that we have a training set $\{x_1, \ldots, x_n\}$ consisting of n independent samples.
  - These samples have some latent variable $\{z_1, \ldots, z_n\}$ jointly distributed with each sample. (For simplicity, let z be discrete)
    - Example. x: image of a digit $\mathbb{R}^{28\times 28}$
      z: digit itself $\{0,1,\ldots,9\}$

Setup

- Suppose that we have a training set $\{x_1, \ldots, x_n\}$ consisting of n independent samples.
  - These samples have some latent variable $\{z_1, \ldots, z_n\}$ jointly distributed with each sample. (For simplicity, let z be discrete)
    - Example. x: image of a digit $\mathbb{R}^{28\times 28}$
      z: digit itself $\{0,1,\ldots,9\}$
- **Goal.** Want to fit a (parametrized) density function

$$p(x;\theta)$$

  - Can be obtained by marginalizing over latent variables

$$p(x;\theta) = \sum_z p(x,z;\theta)$$

Setup

- More specifically, we maximize the log-likelihood of the data:

$$\max_{\theta} \quad \ell(\theta) := \sum_{i=1}^{n} \log p(x_i; \theta)$$

  - In terms of the latent variables, we can write as

$$\ell(\theta) = \sum_{i=1}^{n} \log \sum_{z_i} p(x_i, z_i; \theta)$$

Setup

- More specifically, we maximize the log-likelihood of the data:

$$\max_{\theta} \quad \ell(\theta) := \sum_{i=1}^{n} \log p(x_i; \theta)$$

  - In terms of the latent variables, we can write as

$$\ell(\theta) = \sum_{i=1}^{n} \log \sum_{z_i} p(x_i, z_i; \theta)$$

    - Often, importantly, if z_i were observed then the MLE would have been much easier (e.g., by admitting closed-form solutions)

$$\ell(\theta) = \sum_{i=1}^{n} \log p(x_i, z_i; \theta)$$

Strategy

- **Idea.** Repeat the following:
  - Construct some lower bound on $\tilde{\ell}(\theta) \leq \ell(\theta)$
  - Maximize the lower bound $\tilde{\ell}(\theta)$

Strategy

- **Idea.** Repeat the following:
  - Construct some lower bound on $\tilde{\ell}(\theta) \leq \ell(\theta)$
  - Maximize the lower bound $\tilde{\ell}(\theta)$
- **Simplification.** For simple notation, make it a problem with respect to a single sample

$$\ell(\theta) = \log p(x;\theta) = \log \sum_z p(x,z;\theta)$$

Strategy

- **Idea.** Repeat the following:
  - Construct some lower bound on $\tilde{\ell}(\theta) \leq \ell(\theta)$
  - Maximize the lower bound $\tilde{\ell}(\theta)$
- **Simplification.** For simple notation, make it a problem with respect to a single sample

$$\ell(\theta) = \log p(x;\theta) = \log \sum_z p(x,z;\theta)$$

  - Observation. If we select any distribution $Q(z)$, Jensen's inequality gives

$$\log \sum_z p(x,z;\theta) = \log \sum_z Q(z)\frac{p(x,z;\theta)}{Q(z)} \geq \sum_z Q(z)\log \frac{p(x,z;\theta)}{Q(z)}$$

Strategy

- By letting

$$\tilde{\ell}_Q(\theta) = \sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)}$$

  we know that we have a lower bound

$$\ell(\theta) \geq \tilde{\ell}_Q(\theta)$$

- **Question.** How do we select the tightest lower bound?

Strategy

- By letting

$$\tilde{\ell}_Q(\theta) = \sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)}$$

we know that we have a lower bound

$$\ell(\theta) \geq \tilde{\ell}_Q(\theta)$$

- **Question.** How do we select the tightest lower bound?

    - Answer. We desire that the random quantity is actually constant (equal to "expectation"), i.e.,

$$\frac{p(x, z; \theta)}{Q(z)} = C \qquad \Leftrightarrow \qquad Q(z) \propto p(x, z; \theta)$$

Strategy

- That is, we select $Q(z)$ to be the posterior distribution

$$Q(z) = \frac{p(x, z; \theta)}{p(x; \theta)} = p(z \,|\, x; \theta)$$

  - We call these lower bounds, the **ELBO (evidence lower bound)**

$$\mathrm{ELBO}(x; Q, \theta) = \sum_{z} Q(z) \log \frac{p(x, z; \theta)}{Q(z)} = \tilde{\ell}(\theta) \leq \ell(\theta)$$

  - Remark. Constructing such Q requires some estimate of θ.

Strategy

- That is, we select $Q(z)$ to be the posterior distribution

$$Q(z) = \frac{p(x, z; \theta)}{p(x; \theta)} = p(z \mid x; \theta)$$

  - We call these lower bounds, the **ELBO (evidence lower bound)**

$$\text{ELBO}(x; Q, \theta) = \sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)} = \tilde{\ell}(\theta) \leq \ell(\theta)$$

  - Remark. Constructing such Q requires some estimate of θ.

- We construct such posterior for each sample, i.e.,

$$Q_i(z) = p(z_i \mid x_i, \theta)$$

Strategy

- Thus, our algorithm is repeating:
  - Expectation. Construct some $Q_i(z) = p(z \mid x_i; \theta_{\text{cur}})$ based on the current estimate θ_{cur}
  - Maximization. Find θ_{new} that maximizes

$$\tilde{\ell}(\theta_{\text{new}}) = \sum_{i=1}^{n} \sum_{z} Q_i(z) \log \frac{p(x_i, z_i; \theta_{\text{new}})}{Q_i(z)}$$

Convergence

- To prove convergence, we show that our iteration always improves $\ell(\theta)$, i.e.,

$$\ell(\theta_{\text{cur}}) \leq \ell(\theta_{\text{new}})$$

- In fact, for the distribution Q constructed based on θ_{cur}, we have

$$\begin{aligned} \ell(\theta_{\text{cur}}) &= \text{ELBO}(x; Q, \theta_{\text{cur}}) \\ &\leq \text{ELBO}(x; Q, \theta_{\text{new}}) \\ &\leq \ell(\theta_{\text{new}}) \end{aligned}$$

Next lecture

- Dimensionality reduction

Cheers