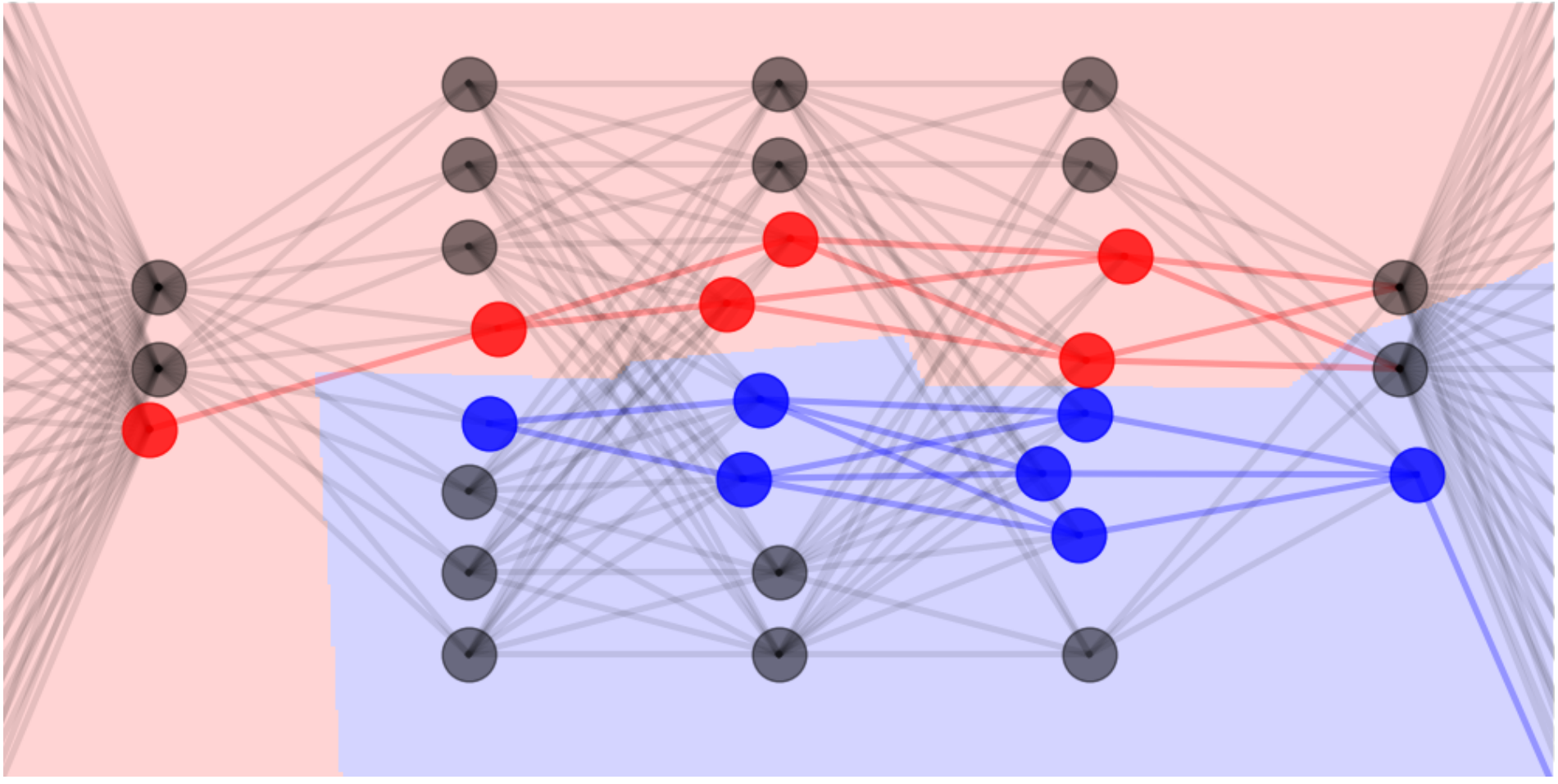
# TU-Delft Machine and Deep Learning



Loss functions

Lecture 2

Brightspace: MDL-Loss-RGrosse.pdf; DL-book: 3,5,6; UDL-book: 5.

Lecturer: Jan van Gemert
Several slides credit to Roger Grosse

# Questions?

Last time:

- Feed forward networks
- Stochastic gradient descent
- XOR

# Questions?

Last time:

- Feed forward networks
- Stochastic gradient descent
- XOR

After this lecture you can:

- Understand maximum likelihood
- Explain the relation between a loss and gradient descent
- Understand binary classification with logistic regression
- Understand multiclass logistic regression

Brightspace: MDL-Loss-RGrosse.pdf
DL book chapters: 3.10, 3.13, 5.5, 5.7.1, 6.2
UDL book chapter: 5

# Maximum likelihood estimation

DL-book: Chapter 5.5; UDL-book: 5.1.

- Training set of $m$ i.i.d. samples $\mathbb{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ from $p_{\text{data}}(x)$

1

# Maximum likelihood estimation

DL-book: Chapter 5.5; UDL-book: 5.1.

- Training set of $m$ i.i.d. samples $\mathbb{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ from $p_{\text{data}}(x)$
- Our model: Family of probability distributions $p_{\text{model}}(x; \theta)$ parameterized by $\theta$

1

# Maximum likelihood estimation

DL-book: Chapter 5.5; UDL-book: 5.1.

- Training set of $m$ i.i.d. samples $\mathbb{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ from $p_{\text{data}}(x)$
- Our model: Family of probability distributions $p_{\text{model}}(x; \theta)$ parameterized by $\theta$
- Q: What is this $p_{\text{model}}(x; \theta)$ a probability over?

1

# Maximum likelihood estimation

DL-book: Chapter 5.5; UDL-book: 5.1.

- Training set of $m$ i.i.d. samples $\mathbb{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ from $p_{\text{data}}(x)$
- Our model: Family of probability distributions $p_{\text{model}}(x; \theta)$ parameterized by $\theta$
- Q: What is this $p_{\text{model}}(x; \theta)$ a probability over? A: all possible parameters.
- Maximum likelihood is $\theta_{ML} = \arg\max_{\theta} p_{\text{model}}(\mathbb{X}; \theta)$

1

# Maximum likelihood estimation

DL-book: Chapter 5.5; UDL-book: 5.1.

- Training set of $m$ i.i.d. samples $\mathbb{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ from $p_{\text{data}}(x)$
- Our model: Family of probability distributions $p_{\text{model}}(x; \theta)$ parameterized by $\theta$
- Q: What is this $p_{\text{model}}(x; \theta)$ a probability over? A: all possible parameters.
- Maximum likelihood is $\theta_{ML} = \arg\max_\theta p_{\text{model}}(\mathbb{X}; \theta)$
- Rewrite in terms of data samples: $\arg\max_\theta \prod_{i=1}^{m} p_{\text{model}}(x^{(i)}; \theta)$

1

# Maximum likelihood estimation

DL-book: Chapter 5.5; UDL-book: 5.1.

- Training set of $m$ i.i.d. samples $\mathbb{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ from $p_{\text{data}}(x)$
- Our model: Family of probability distributions $p_{\text{model}}(x; \theta)$ parameterized by $\theta$
- Q: What is this $p_{\text{model}}(x; \theta)$ a probability over? A: all possible parameters.
- Maximum likelihood is $\theta_{ML} = \arg\max_\theta p_{\text{model}}(\mathbb{X}; \theta)$
- Rewrite in terms of data samples: $\arg\max_\theta \prod_{i=1}^{m} p_{\text{model}}(x^{(i)}; \theta)$
- Q: What assumption allowed this?

1

# Maximum likelihood estimation

DL-book: Chapter 5.5; UDL-book: 5.1.

- Training set of $m$ i.i.d. samples $\mathbb{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ from $p_{\text{data}}(x)$
- Our model: Family of probability distributions $p_{\text{model}}(x; \theta)$ parameterized by $\theta$
- Q: What is this $p_{\text{model}}(x; \theta)$ a probability over? A: all possible parameters.
- Maximum likelihood is $\theta_{ML} = \arg\max_\theta p_{\text{model}}(\mathbb{X}; \theta)$
- Rewrite in terms of data samples: $\arg\max_\theta \prod_{i=1}^{m} p_{\text{model}}(x^{(i)}; \theta)$
- Q: What assumption allowed this? A: i.i.d.

1

# Maximum likelihood estimation

DL-book: Chapter 5.5; UDL-book: 5.1.

- Training set of $m$ i.i.d. samples $\mathbb{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ from $p_{\text{data}}(x)$
- Our model: Family of probability distributions $p_{\text{model}}(x; \theta)$ parameterized by $\theta$
- Q: What is this $p_{\text{model}}(x; \theta)$ a probability over? A: all possible parameters.
- Maximum likelihood is $\theta_{ML} = \arg\max_\theta p_{\text{model}}(\mathbb{X}; \theta)$
- Rewrite in terms of data samples: $\arg\max_\theta \prod_{i=1}^{m} p_{\text{model}}(x^{(i)}; \theta)$
- Q: What assumption allowed this? A: i.i.d.
- Q: What's a computational problem with multiplying values between 0 and 1?

1

# Maximum likelihood estimation

DL-book: Chapter 5.5; UDL-book: 5.1.

- Training set of $m$ i.i.d. samples $\mathbb{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ from $p_{\text{data}}(x)$
- Our model: Family of probability distributions $p_{\text{model}}(x; \theta)$ parameterized by $\theta$
- Q: What is this $p_{\text{model}}(x; \theta)$ a probability over? A: all possible parameters.
- Maximum likelihood is $\theta_{ML} = \arg\max_\theta p_{\text{model}}(\mathbb{X}; \theta)$
- Rewrite in terms of data samples: $\arg\max_\theta \prod_{i=1}^{m} p_{\text{model}}(x^{(i)}; \theta)$
- Q: What assumption allowed this? A: i.i.d.
- Q: What's a computational problem with multiplying values between 0 and 1? A: Numerically unstable when close to 0.

[1] The $\hat{p}$ instead of $p$ is used when $p$ is sampled (ie: *empirical* means sampled by data)

# Maximum likelihood estimation

DL-book: Chapter 5.5; UDL-book: 5.1.

- Training set of $m$ i.i.d. samples $\mathbb{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ from $p_{\text{data}}(x)$
- Our model: Family of probability distributions $p_{\text{model}}(x; \theta)$ parameterized by $\theta$
- Q: What is this $p_{\text{model}}(x; \theta)$ a probability over? A: all possible parameters.
- Maximum likelihood is $\theta_{ML} = \arg\max_\theta p_{\text{model}}(\mathbb{X}; \theta)$
- Rewrite in terms of data samples: $\arg\max_\theta \prod_{i=1}^m p_{\text{model}}(x^{(i)}; \theta)$
- Q: What assumption allowed this? A: i.i.d.
- Q: What's a computational problem with multiplying values between 0 and 1? A: Numerically unstable when close to 0.
- Log: same arg-max, turns multiplications to sums:
  $\arg\max_\theta \sum_{i=1}^m \log p_{\text{model}}(x^{(i)}; \theta)$

[1] The $\hat{p}$ instead of $p$ is used when $p$ is sampled (ie: *empirical* means sampled by data)

# Maximum likelihood estimation

DL-book: Chapter 5.5; UDL-book: 5.1.

- Training set of $m$ i.i.d. samples $\mathbb{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ from $p_{\text{data}}(x)$
- Our model: Family of probability distributions $p_{\text{model}}(x; \theta)$ parameterized by $\theta$
- Q: What is this $p_{\text{model}}(x; \theta)$ a probability over? A: all possible parameters.
- Maximum likelihood is $\theta_{ML} = \arg\max_\theta p_{\text{model}}(\mathbb{X}; \theta)$
- Rewrite in terms of data samples: $\arg\max_\theta \prod_{i=1}^m p_{\text{model}}(x^{(i)}; \theta)$
- Q: What assumption allowed this? A: i.i.d.
- Q: What's a computational problem with multiplying values between 0 and 1? A: Numerically unstable when close to 0.
- Log: same arg-max, turns multiplications to sums:
  $\arg\max_\theta \sum_{i=1}^m \log p_{\text{model}}(x^{(i)}; \theta)$
- Rewrite sum as *expectation* wrt empirical distribution[1]
  $\hat{p}_{\text{data}}$ as $\arg\max_\theta \mathbb{E}_{x \sim \hat{p}_{\text{data}}} \log p_{\text{model}}(x; \theta)$,

[1] The $\hat{p}$ instead of $p$ is used when $p$ is sampled (ie: *empirical* means sampled by data)

# Maximum likelihood estimation

DL-book: Chapter 5.5; UDL-book: 5.1.

- Training set of $m$ i.i.d. samples $\mathbb{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ from $p_{\text{data}}(x)$
- Our model: Family of probability distributions $p_{\text{model}}(x; \theta)$ parameterized by $\theta$
- Q: What is this $p_{\text{model}}(x; \theta)$ a probability over? A: all possible parameters.
- Maximum likelihood is $\theta_{ML} = \arg\max_\theta p_{\text{model}}(\mathbb{X}; \theta)$
- Rewrite in terms of data samples: $\arg\max_\theta \prod_{i=1}^m p_{\text{model}}(x^{(i)}; \theta)$
- Q: What assumption allowed this? A: i.i.d.
- Q: What's a computational problem with multiplying values between 0 and 1? A: Numerically unstable when close to 0.
- Log: same arg-max, turns multiplications to sums:
  $\arg\max_\theta \sum_{i=1}^m \log p_{\text{model}}(x^{(i)}; \theta)$
- Rewrite sum as *expectation* wrt empirical distribution[1]
  $\hat{p}_{\text{data}}$ as $\arg\max_\theta \mathbb{E}_{x \sim \hat{p}_{\text{data}}} \log p_{\text{model}}(x; \theta)$,
- Q: Why is this the same?

[1] The $\hat{p}$ instead of $p$ is used when $p$ is sampled (ie: *empirical* means sampled by data)

# Maximum likelihood estimation

DL-book: Chapter 5.5; UDL-book: 5.1.

- Training set of $m$ i.i.d. samples $\mathbb{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ from $p_{\text{data}}(x)$
- Our model: Family of probability distributions $p_{\text{model}}(x; \theta)$ parameterized by $\theta$
- Q: What is this $p_{\text{model}}(x; \theta)$ a probability over? A: all possible parameters.
- Maximum likelihood is $\theta_{ML} = \arg\max_\theta p_{\text{model}}(\mathbb{X}; \theta)$
- Rewrite in terms of data samples: $\arg\max_\theta \prod_{i=1}^m p_{\text{model}}(x^{(i)}; \theta)$
- Q: What assumption allowed this? A: i.i.d.
- Q: What's a computational problem with multiplying values between 0 and 1? A: Numerically unstable when close to 0.
- Log: same arg-max, turns multiplications to sums:
  $\arg\max_\theta \sum_{i=1}^m \log p_{\text{model}}(x^{(i)}; \theta)$
- Rewrite sum as *expectation* wrt empirical distribution[1]
  $\hat{p}_{\text{data}}$ as $\arg\max_\theta \mathbb{E}_{x \sim \hat{p}_{\text{data}}} \log p_{\text{model}}(x; \theta)$,
- Q: Why is this the same? A: Dividing by $m$ does not change the arg-max

[1] The $\hat{p}$ instead of $p$ is used when $p$ is sampled (ie: *empirical* means sampled by data)

# Minimize the dissimilarity of distributions

DL book: 5.5, 3.13; UDL-book: 5.7.

- Max likelihood: $\arg\max_\theta \mathbb{E}_{x\sim\hat{p}_{\text{data}}} \log p_{\text{model}}(x;\theta)$
- One interpretation: Minimize dissimilarity between $\hat{p}_{\text{data}}$ and $p_{\text{model}}$

# Minimize the dissimilarity of distributions

DL book: 5.5, 3.13; UDL-book: 5.7.

- Max likelihood: $\arg\max_\theta \mathbb{E}_{x\sim\hat{p}_{\text{data}}} \log p_{\text{model}}(x;\theta)$
- One interpretation: Minimize dissimilarity between $\hat{p}_{\text{data}}$ and $p_{\text{model}}$
- Measure dissimilarity by the KL divergence $D$

# Minimize the dissimilarity of distributions

DL book: 5.5, 3.13; UDL-book: 5.7.

- Max likelihood: $\arg\max_\theta \mathbb{E}_{x\sim\hat{p}_{\text{data}}} \log p_{\text{model}}(x;\theta)$
- One interpretation: Minimize dissimilarity between $\hat{p}_{\text{data}}$ and $p_{\text{model}}$
- Measure dissimilarity by the KL divergence $D$

$$D(p_{\text{data}}||p_{\text{model}}) = \sum_{i=1}^{m} \hat{p}_{\text{data}}(x^{(i)}) \log\left(\frac{\hat{p}_{\text{data}}(x^{(i)})}{p_{\text{model}}(x^{(i)};\theta)}\right)$$

# Minimize the dissimilarity of distributions

DL book: 5.5, 3.13; UDL-book: 5.7.

- Max likelihood: $\arg\max_\theta \mathbb{E}_{x\sim\hat{p}_{\text{data}}} \log p_{\text{model}}(x;\theta)$
- One interpretation: Minimize dissimilarity between $\hat{p}_{\text{data}}$ and $p_{\text{model}}$
- Measure dissimilarity by the KL divergence $D$

$$D(p_{\text{data}}||p_{\text{model}}) = \sum_{i=1}^{m} \hat{p}_{\text{data}}(x^{(i)}) \log\left(\frac{\hat{p}_{\text{data}}(x^{(i)})}{p_{\text{model}}(x^{(i)};\theta)}\right)$$

- Q: It uses a log, how to rewrite the division?

# Minimize the dissimilarity of distributions

DL book: 5.5, 3.13; UDL-book: 5.7.

- Max likelihood: $\arg\max_\theta \mathbb{E}_{x\sim\hat{p}_{\text{data}}} \log p_{\text{model}}(x;\theta)$
- One interpretation: Minimize dissimilarity between $\hat{p}_{\text{data}}$ and $p_{\text{model}}$
- Measure dissimilarity by the KL divergence $D$

$$D(p_{\text{data}}||p_{\text{model}}) = \sum_{i=1}^{m} \hat{p}_{\text{data}}(x^{(i)}) \log\left(\frac{\hat{p}_{\text{data}}(x^{(i)})}{p_{\text{model}}(x^{(i)};\theta)}\right)$$

- Q: It uses a log, how to rewrite the division?
  A: $\sum_{i=1}^{m} \hat{p}_{\text{data}}(x^{(i)}) \left(\log \hat{p}_{\text{data}}(x^{(i)}) - \log p_{\text{model}}(x^{(i)};\theta)\right)$
- Q: rewrite as expectation over data samples $\hat{p}_{\text{data}}$?

# Minimize the dissimilarity of distributions

DL book: 5.5, 3.13; UDL-book: 5.7.

- Max likelihood: $\arg\max_\theta \mathbb{E}_{x\sim\hat{p}_{\text{data}}} \log p_{\text{model}}(x;\theta)$
- One interpretation: Minimize dissimilarity between $\hat{p}_{\text{data}}$ and $p_{\text{model}}$
- Measure dissimilarity by the KL divergence $D$

$$D(p_{\text{data}}||p_{\text{model}}) = \sum_{i=1}^{m} \hat{p}_{\text{data}}(x^{(i)}) \log\left(\frac{\hat{p}_{\text{data}}(x^{(i)})}{p_{\text{model}}(x^{(i)};\theta)}\right)$$

- Q: It uses a log, how to rewrite the division?
  A: $\sum_{i=1}^{m} \hat{p}_{\text{data}}(x^{(i)}) \left(\log \hat{p}_{\text{data}}(x^{(i)}) - \log p_{\text{model}}(x^{(i)};\theta)\right)$
- Q: rewrite as expectation over data samples $\hat{p}_{\text{data}}$?
  A: $\mathbb{E}_{x\sim\hat{p}_{\text{data}}} \left[\log \hat{p}_{\text{data}}(x) - \log p_{\text{model}}(x;\theta)\right]$

# Minimize the dissimilarity of distributions

DL book: 5.5, 3.13; UDL-book: 5.7.

- Max likelihood: $\arg\max_\theta \mathbb{E}_{x\sim\hat{p}_{\text{data}}} \log p_{\text{model}}(x;\theta)$
- One interpretation: Minimize dissimilarity between $\hat{p}_{\text{data}}$ and $p_{\text{model}}$
- Measure dissimilarity by the KL divergence $D$

$$D(p_{\text{data}}||p_{\text{model}}) = \sum_{i=1}^{m} \hat{p}_{\text{data}}(x^{(i)}) \log\left(\frac{\hat{p}_{\text{data}}(x^{(i)})}{p_{\text{model}}(x^{(i)};\theta)}\right)$$

- Q: It uses a log, how to rewrite the division?
  A: $\sum_{i=1}^{m} \hat{p}_{\text{data}}(x^{(i)}) \left(\log \hat{p}_{\text{data}}(x^{(i)}) - \log p_{\text{model}}(x^{(i)};\theta)\right)$
- Q: rewrite as expectation over data samples $\hat{p}_{\text{data}}$?
  A: $\mathbb{E}_{x\sim\hat{p}_{\text{data}}} \left[\log \hat{p}_{\text{data}}(x) - \log p_{\text{model}}(x;\theta)\right]$
- Q: where did the $\hat{p}_{\text{data}}(x^{(i)})$ go?

# Minimize the dissimilarity of distributions

DL book: 5.5, 3.13; UDL-book: 5.7.

- Max likelihood: $\arg\max_\theta \mathbb{E}_{x\sim\hat{p}_{\text{data}}} \log p_{\text{model}}(x;\theta)$
- One interpretation: Minimize dissimilarity between $\hat{p}_{\text{data}}$ and $p_{\text{model}}$
- Measure dissimilarity by the KL divergence $D$

$$D(p_{\text{data}}||p_{\text{model}}) = \sum_{i=1}^{m} \hat{p}_{\text{data}}(x^{(i)}) \log\left(\frac{\hat{p}_{\text{data}}(x^{(i)})}{p_{\text{model}}(x^{(i)};\theta)}\right)$$

- Q: It uses a log, how to rewrite the division?
  A: $\sum_{i=1}^{m} \hat{p}_{\text{data}}(x^{(i)}) \left(\log \hat{p}_{\text{data}}(x^{(i)}) - \log p_{\text{model}}(x^{(i)};\theta)\right)$
- Q: rewrite as expectation over data samples $\hat{p}_{\text{data}}$?
  A: $\mathbb{E}_{x\sim\hat{p}_{\text{data}}} \left[\log \hat{p}_{\text{data}}(x) - \log p_{\text{model}}(x;\theta)\right]$
- Q: where did the $\hat{p}_{\text{data}}(x^{(i)})$ go?
  A: Assume each data point is equally likely: $\frac{1}{m}$

# Minimize the dissimilarity of distributions

DL book: 5.5, 3.13; UDL-book: 5.7.

- Max likelihood: $\arg\max_\theta \mathbb{E}_{x\sim\hat{p}_{\text{data}}} \log p_{\text{model}}(x;\theta)$
- One interpretation: Minimize dissimilarity between $\hat{p}_{\text{data}}$ and $p_{\text{model}}$
- Measure dissimilarity by the KL divergence $D$

$$D(p_{\text{data}}||p_{\text{model}}) = \sum_{i=1}^{m} \hat{p}_{\text{data}}(x^{(i)}) \log\left(\frac{\hat{p}_{\text{data}}(x^{(i)})}{p_{\text{model}}(x^{(i)};\theta)}\right)$$

- Q: It uses a log, how to rewrite the division?
  A: $\sum_{i=1}^{m} \hat{p}_{\text{data}}(x^{(i)}) \left(\log \hat{p}_{\text{data}}(x^{(i)}) - \log p_{\text{model}}(x^{(i)};\theta)\right)$
- Q: rewrite as expectation over data samples $\hat{p}_{\text{data}}$?
  A: $\mathbb{E}_{x\sim\hat{p}_{\text{data}}} \left[\log \hat{p}_{\text{data}}(x) - \log p_{\text{model}}(x;\theta)\right]$
- Q: where did the $\hat{p}_{\text{data}}(x^{(i)})$ go?
  A: Assume each data point is equally likely: $\frac{1}{m}$
- For the best model, left/data term $(\log \hat{p}_{\text{data}}(x))$ does not matter: Remove.

# Minimize the dissimilarity of distributions

DL book: 5.5, 3.13; UDL-book: 5.7.

- Max likelihood: $\arg\max_\theta \mathbb{E}_{x\sim\hat{p}_{\text{data}}} \log p_{\text{model}}(x;\theta)$
- One interpretation: Minimize dissimilarity between $\hat{p}_{\text{data}}$ and $p_{\text{model}}$
- Measure dissimilarity by the KL divergence $D$

$$D(p_{\text{data}}||p_{\text{model}}) = \sum_{i=1}^{m} \hat{p}_{\text{data}}(x^{(i)}) \log\left(\frac{\hat{p}_{\text{data}}(x^{(i)})}{p_{\text{model}}(x^{(i)};\theta)}\right)$$

- Q: It uses a log, how to rewrite the division?
  A: $\sum_{i=1}^{m} \hat{p}_{\text{data}}(x^{(i)}) \left(\log \hat{p}_{\text{data}}(x^{(i)}) - \log p_{\text{model}}(x^{(i)};\theta)\right)$
- Q: rewrite as expectation over data samples $\hat{p}_{\text{data}}$?
  A: $\mathbb{E}_{x\sim\hat{p}_{\text{data}}} [\log \hat{p}_{\text{data}}(x) - \log p_{\text{model}}(x;\theta)]$
- Q: where did the $\hat{p}_{\text{data}}(x^{(i)})$ go?
  A: Assume each data point is equally likely: $\frac{1}{m}$
- For the best model, left/data term $(\log \hat{p}_{\text{data}}(x))$ does not matter: Remove.
- So, minimize $\mathbb{E}_{x\sim\hat{p}_{\text{data}}} [-\log p_{\text{model}}(x;\theta)]$

# Minimize the dissimilarity of distributions

DL book: 5.5, 3.13; UDL-book: 5.7.

- Max likelihood: $\arg\max_\theta \mathbb{E}_{x\sim\hat{p}_{\text{data}}} \log p_{\text{model}}(x;\theta)$
- One interpretation: Minimize dissimilarity between $\hat{p}_{\text{data}}$ and $p_{\text{model}}$
- Measure dissimilarity by the KL divergence $D$

$$D(p_{\text{data}}||p_{\text{model}}) = \sum_{i=1}^{m} \hat{p}_{\text{data}}(x^{(i)}) \log\left(\frac{\hat{p}_{\text{data}}(x^{(i)})}{p_{\text{model}}(x^{(i)};\theta)}\right)$$

- Q: It uses a log, how to rewrite the division?
  A: $\sum_{i=1}^{m} \hat{p}_{\text{data}}(x^{(i)}) \left(\log \hat{p}_{\text{data}}(x^{(i)}) - \log p_{\text{model}}(x^{(i)};\theta)\right)$
- Q: rewrite as expectation over data samples $\hat{p}_{\text{data}}$?
  A: $\mathbb{E}_{x\sim\hat{p}_{\text{data}}} [\log \hat{p}_{\text{data}}(x) - \log p_{\text{model}}(x;\theta)]$
- Q: where did the $\hat{p}_{\text{data}}(x^{(i)})$ go?
  A: Assume each data point is equally likely: $\frac{1}{m}$
- For the best model, left/data term $(\log \hat{p}_{\text{data}}(x))$ does not matter: Remove.
- So, minimize $\mathbb{E}_{x\sim\hat{p}_{\text{data}}} [-\log p_{\text{model}}(x;\theta)]$
- Q: Relate $\arg\min_\theta -\mathbb{E}_{x\sim\hat{p}_{\text{data}}} [\log p_{\text{model}}(x;\theta)]$ to max likelihood?

# Minimize the dissimilarity of distributions

DL book: 5.5, 3.13; UDL-book: 5.7.

- Max likelihood: $\arg\max_\theta \mathbb{E}_{x\sim\hat{p}_{\text{data}}} \log p_{\text{model}}(x;\theta)$
- One interpretation: Minimize dissimilarity between $\hat{p}_{\text{data}}$ and $p_{\text{model}}$
- Measure dissimilarity by the KL divergence $D$

$$D(p_{\text{data}}||p_{\text{model}}) = \sum_{i=1}^{m} \hat{p}_{\text{data}}(x^{(i)}) \log\left(\frac{\hat{p}_{\text{data}}(x^{(i)})}{p_{\text{model}}(x^{(i)};\theta)}\right)$$

- Q: It uses a log, how to rewrite the division?
  A: $\sum_{i=1}^{m} \hat{p}_{\text{data}}(x^{(i)}) \left(\log \hat{p}_{\text{data}}(x^{(i)}) - \log p_{\text{model}}(x^{(i)};\theta)\right)$
- Q: rewrite as expectation over data samples $\hat{p}_{\text{data}}$?
  A: $\mathbb{E}_{x\sim\hat{p}_{\text{data}}} [\log \hat{p}_{\text{data}}(x) - \log p_{\text{model}}(x;\theta)]$
- Q: where did the $\hat{p}_{\text{data}}(x^{(i)})$ go?
  A: Assume each data point is equally likely: $\frac{1}{m}$
- For the best model, left/data term $(\log \hat{p}_{\text{data}}(x))$ does not matter: Remove.
- So, minimize $\mathbb{E}_{x\sim\hat{p}_{\text{data}}} [-\log p_{\text{model}}(x;\theta)]$
- Q: Relate $\arg\min_\theta -\mathbb{E}_{x\sim\hat{p}_{\text{data}}} [\log p_{\text{model}}(x;\theta)]$ to max likelihood? A: Same.

# Cross-entropy

DL-book: 5.5, 3.13; UDL-book: 5.7.

The literature optimizes "cross-entropy".

- KL related to cross-entropy between two distributions $H(p_{\text{data}}, p_{\text{model}}) = H(p_{\text{data}}) + D(p_{\text{data}}||p_{\text{model}})$, where $H$ is entropy.

# Cross-entropy

DL-book: 5.5, 3.13; UDL-book: 5.7.

The literature optimizes "cross-entropy".

- KL related to cross-entropy between two distributions $H(p_{\text{data}}, p_{\text{model}}) = H(p_{\text{data}}) + D(p_{\text{data}}||p_{\text{model}})$, where $H$ is entropy.
- Q: What is the effect of $H(p_{\text{data}})$ on $p_{\text{model}}$?

# Cross-entropy

DL-book: 5.5, 3.13; UDL-book: 5.7.

The literature optimizes "cross-entropy".

- KL related to cross-entropy between two distributions $H(p_{\text{data}}, p_{\text{model}}) = H(p_{\text{data}}) + D(p_{\text{data}}||p_{\text{model}})$, where $H$ is entropy.
- Q: What is the effect of $H(p_{\text{data}})$ on $p_{\text{model}}$?
  A: The model does not depend on $H(p_{\text{data}})$, so can be removed.

# Cross-entropy

DL-book: 5.5, 3.13; UDL-book: 5.7.

The literature optimizes "cross-entropy".

- KL related to cross-entropy between two distributions $H(p_{\text{data}}, p_{\text{model}}) = H(p_{\text{data}}) + D(p_{\text{data}}||p_{\text{model}})$, where $H$ is entropy.
- Q: What is the effect of $H(p_{\text{data}})$ on $p_{\text{model}}$?
  A: The model does not depend on $H(p_{\text{data}})$, so can be removed.
- The term cross-entropy is generic, not just for classification (Bernoulli or softmax distributions).

# Cross-entropy

DL-book: 5.5, 3.13; UDL-book: 5.7.

The literature optimizes "cross-entropy".

- KL related to cross-entropy between two distributions $H(p_{\text{data}}, p_{\text{model}}) = H(p_{\text{data}}) + D(p_{\text{data}}||p_{\text{model}})$, where $H$ is entropy.
- Q: What is the effect of $H(p_{\text{data}})$ on $p_{\text{model}}$?
  A: The model does not depend on $H(p_{\text{data}})$, so can be removed.
- The term cross-entropy is generic, not just for classification (Bernoulli or softmax distributions).
- Minimize KL divergence $\Leftrightarrow$ minimize negative log likelihood $\Leftrightarrow$ minimize cross-entropy $\Leftrightarrow$ maximize maximum likelihood.

# Conditional log-likelihood and output units

DL-book: 5.5.1 and 6.2.2

Often interested in classification problems

- Conditional log-likelihood: Predict labels $Y$ given data $X$: $P(Y|X;\theta)$
- $\theta_{ML} = \arg\max_\theta P(Y|X;\theta)$
- Assume i.i.d.: $\arg\max_\theta \sum_{i=1}^{m} \log P(y^{(i)}|x^{(i)};\theta)$

# Conditional log-likelihood and output units

DL-book: 5.5.1 and 6.2.2

Often interested in classification problems

- Conditional log-likelihood: Predict labels $Y$ given data $X$: $P(Y|X;\theta)$
- $\theta_{ML} = \arg\max_\theta P(Y|X;\theta)$
- Assume i.i.d.: $\arg\max_\theta \sum_{i=1}^{m} \log P(y^{(i)}|x^{(i)};\theta)$

What to optimize in a deep net:

- Usually: cross-entropy between data distribution and model distribution
- Output units determine the form of the cross-entropy function (next topic)

# Questions?

# Binary classification

DL-book: 3.10 and 6.2.2.2

Bernoulli distribution on $P(Y = 1|X)$, single number between $[0, 1]$

# Binary classification

DL-book: 3.10 and 6.2.2.2

Bernoulli distribution on $P(Y = 1|X)$, single number between $[0, 1]$

- Q: Why is a single number enough for two classes?

# Binary classification

DL-book: 3.10 and 6.2.2.2

Bernoulli distribution on $P(Y = 1|X)$, single number between $[0, 1]$

- Q: Why is a single number enough for two classes?
  A: $P(Y = 0|X) = 1 - P(Y = 1|X)$

# Binary classification

DL-book: 3.10 and 6.2.2.2

Bernoulli distribution on $P(Y = 1|X)$, single number between $[0, 1]$

- Q: Why is a single number enough for two classes?
  A: $P(Y = 0|X) = 1 - P(Y = 1|X)$

Lets try linear unit, for input $h$:

$$P(Y = 1|X) = \max\left\{0, \min\left\{1, \mathbf{w}^\top \mathbf{h} + \mathbf{b}\right\}\right\} = y$$

# Binary classification

DL-book: 3.10 and 6.2.2.2

Bernoulli distribution on $P(Y=1|X)$, single number between $[0,1]$

- Q: Why is a single number enough for two classes?
  A: $P(Y=0|X) = 1 - P(Y=1|X)$

Lets try linear unit, for input $h$:

$$P(Y=1|X) = \max\left\{0, \min\left\{1, \mathbf{w}^\top \mathbf{h} + \mathbf{b}\right\}\right\} = y$$

- Q: What is the goal of the min/max terms?

# Binary classification

DL-book: 3.10 and 6.2.2.2

Bernoulli distribution on $P(Y = 1|X)$, single number between $[0, 1]$

- Q: Why is a single number enough for two classes?
  A: $P(Y = 0|X) = 1 - P(Y = 1|X)$

Lets try linear unit, for input $h$:

$$P(Y = 1|X) = \max\left\{0, \min\left\{1, \mathbf{w}^\top \mathbf{h} + \mathbf{b}\right\}\right\} = y$$

- Q: What is the goal of the min/max terms?
  A: To restrict the output between 0 and 1. (a probability)

# Binary classification

DL-book: 3.10 and 6.2.2.2

Bernoulli distribution on $P(Y = 1|X)$, single number between $[0, 1]$

- Q: Why is a single number enough for two classes?
  A: $P(Y = 0|X) = 1 - P(Y = 1|X)$

Lets try linear unit, for input $h$:

$$P(Y = 1|X) = \max\left\{0, \min\left\{1, \mathbf{w}^\top \mathbf{h} + \mathbf{b}\right\}\right\} = y$$

- Q: What is the goal of the min/max terms?
  A: To restrict the output between 0 and 1. (a probability)
- Q: What happens to the gradients outside interval $[0, 1]$?

# Binary classification

DL-book: 3.10 and 6.2.2.2

Bernoulli distribution on $P(Y = 1|X)$, single number between $[0, 1]$

- Q: Why is a single number enough for two classes?
  A: $P(Y = 0|X) = 1 - P(Y = 1|X)$

Lets try linear unit, for input $h$:

$$P(Y = 1|X) = \max\left\{0, \min\left\{1, \mathbf{w}^\top \mathbf{h} + \mathbf{b}\right\}\right\} = y$$

- Q: What is the goal of the min/max terms?
  A: To restrict the output between 0 and 1. (a probability)
- Q: What happens to the gradients outside interval $[0, 1]$?
  A: No more gradients: Cannot be optimized by gradient descent.

# The interval $[0, 1]$

DL-book: 5.7.1, 6.2.2.2, and eq (8) in MDL-Loss-RGrosse.pdf

The min/max prevent gradient flow $P(Y = 1|X) = \max\left\{0, \min\left\{1, w^\top h + b\right\}\right\}$
Lets make them flow again; and remove min/max.

- Q: Other way to limit the output to $[0, 1]$ ?

# The interval $[0, 1]$

DL-book: 5.7.1, 6.2.2.2, and eq (8) in MDL-Loss-RGrosse.pdf

The min/max prevent gradient flow $P(Y = 1|X) = \max\left\{0, \min\left\{1, w^\top h + b\right\}\right\}$
Lets make them flow again; and remove min/max.

- Q: Other way to limit the output to $[0, 1]$ ?
  A: Squash it between 0 and 1.

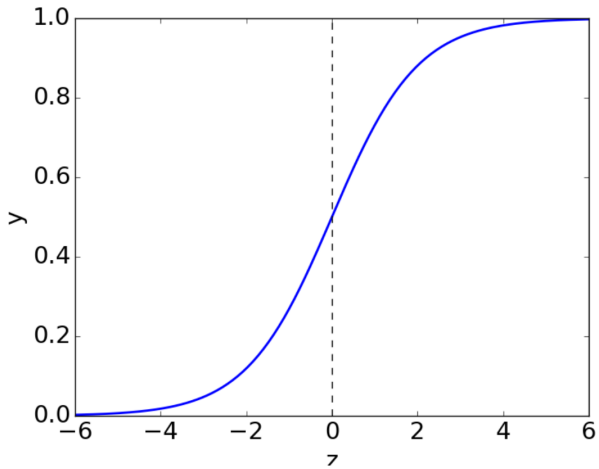
# The interval $[0, 1]$

DL-book: 5.7.1, 6.2.2.2, and eq (8) in MDL-Loss-RGrosse.pdf

The min/max prevent gradient flow $P(Y = 1|X) = \max\left\{0, \min\left\{1, w^\top h + b\right\}\right\}$
Lets make them flow again; and remove min/max.

- Q: Other way to limit the output to $[0, 1]$ ?
  A: Squash it between 0 and 1.
- The logistic function is sigmoidal (S-shaped)

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$



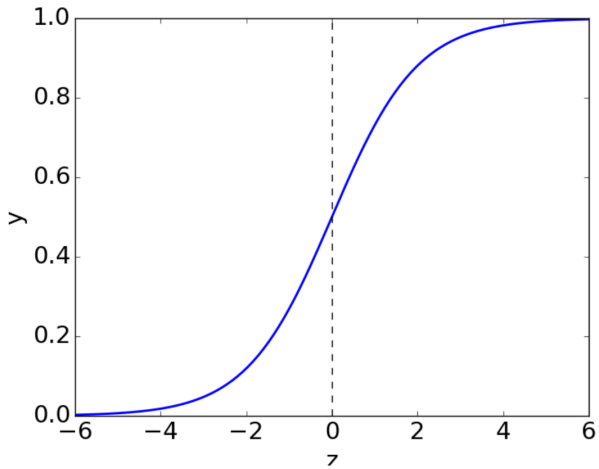
# The interval $[0, 1]$

DL-book: 5.7.1, 6.2.2.2, and eq (8) in MDL-Loss-RGrosse.pdf

The min/max prevent gradient flow $P(Y = 1|X) = \max\left\{0, \min\left\{1, w^\top h + b\right\}\right\}$
Lets make them flow again; and remove min/max.

- Q: Other way to limit the output to $[0, 1]$ ?
  A: Squash it between 0 and 1.
- The logistic function is sigmoidal (S-shaped)

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$




A linear model with a logistic nonlinearity is known as log-linear:

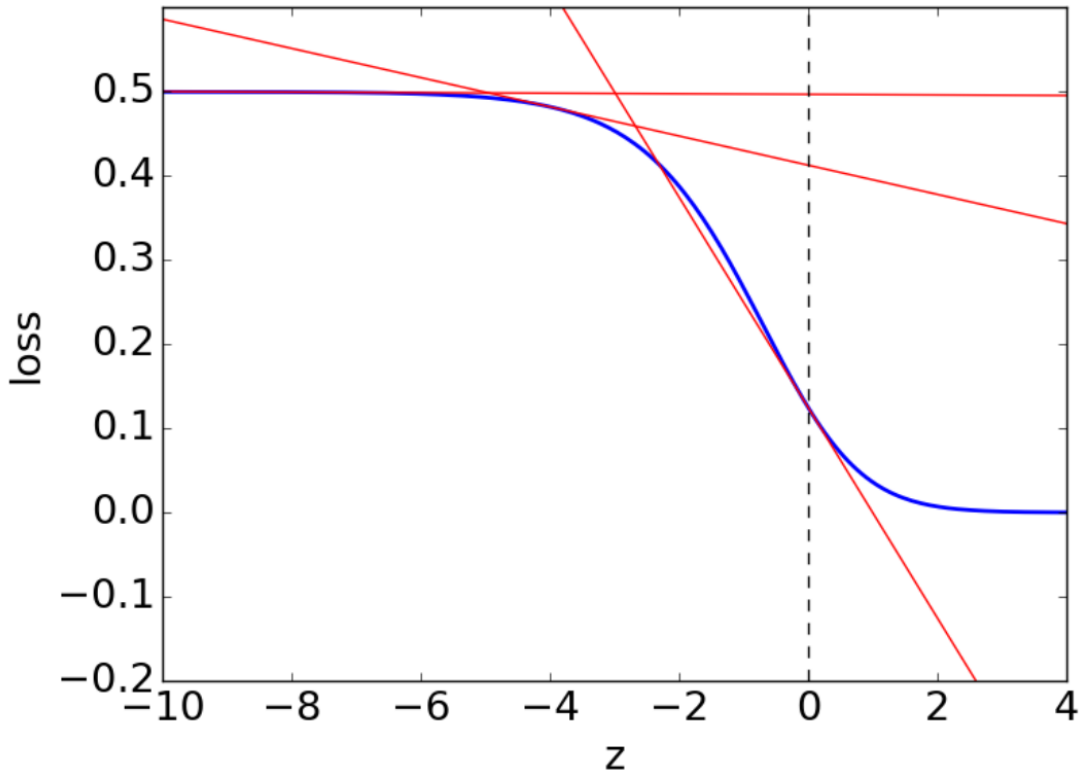
$$z = w^\top x + b$$
$$y = \sigma(z)$$
$$\mathcal{L}_{\text{SE}} = \frac{1}{2}(y - t)^2$$

(SE: Squared Error)

Where $\sigma$ is the activation function and $z$ is called the logit.

# Lets look at the derivatives

Fig 1 in MDL-Loss-RGrosse.pdf

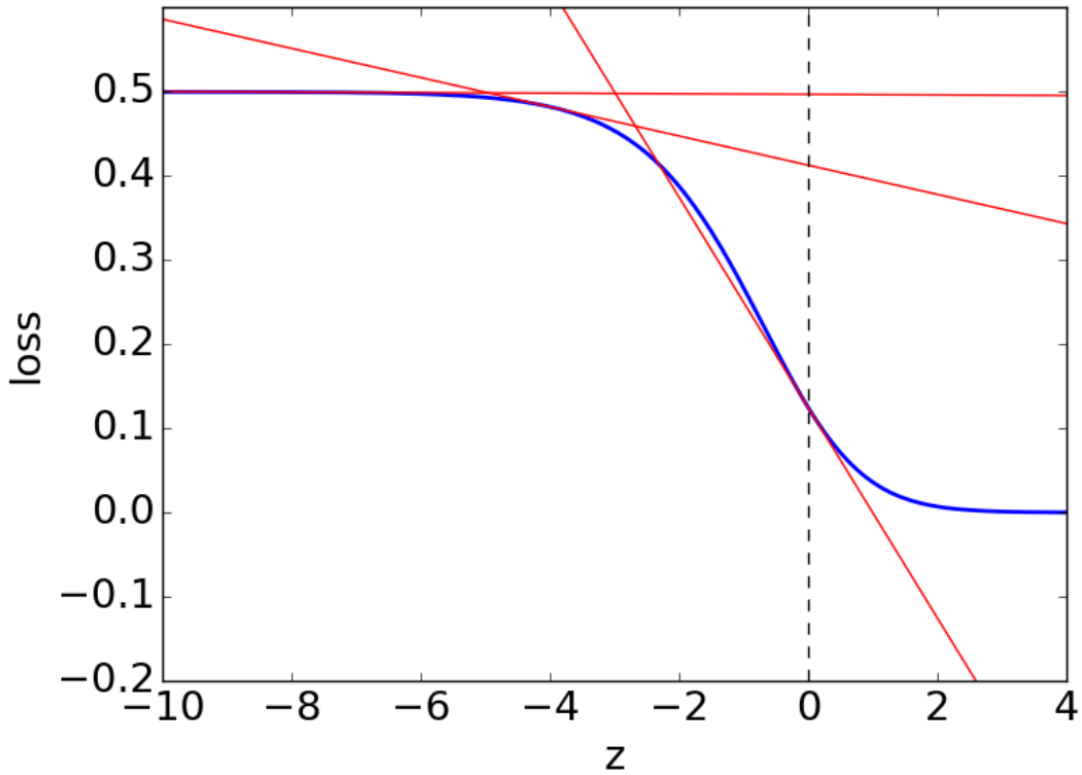



Derivatives of squared error loss with logistic nonlinearity for a **positive** sample.

- Q: Explain what it shows; What is the problem here?

# Lets look at the derivatives

Fig 1 in MDL-Loss-RGrosse.pdf

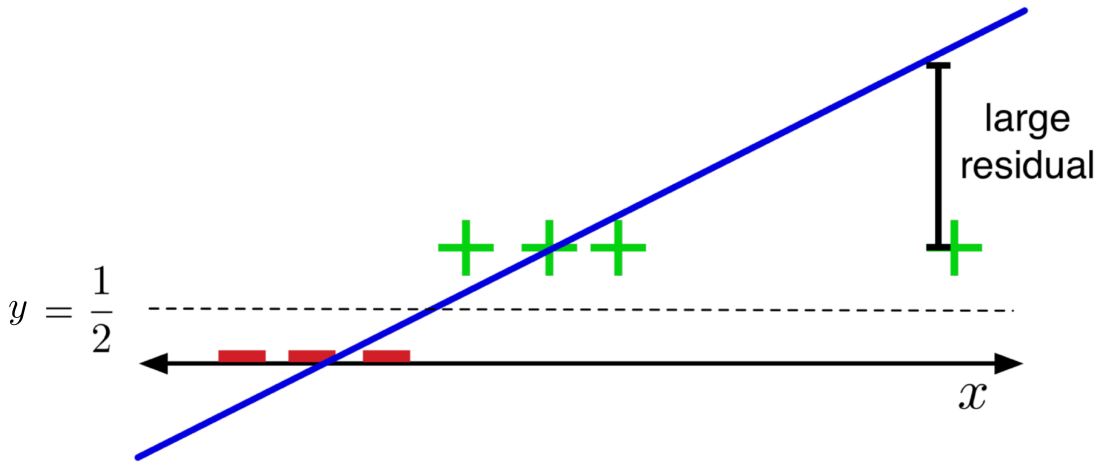



Derivatives of squared error loss with logistic nonlinearity for a **positive** sample.

- Q: Explain what it shows; What is the problem here?
  A: Gradient descent: small gradient is a small step.
- Should take large step when the prediction is really wrong

# loss wants to be exact

The loss used was a square[2] loss: $\mathcal{L}_{\text{SE}} = \frac{1}{2}(y-t)^2$ (SE: Squared Error)

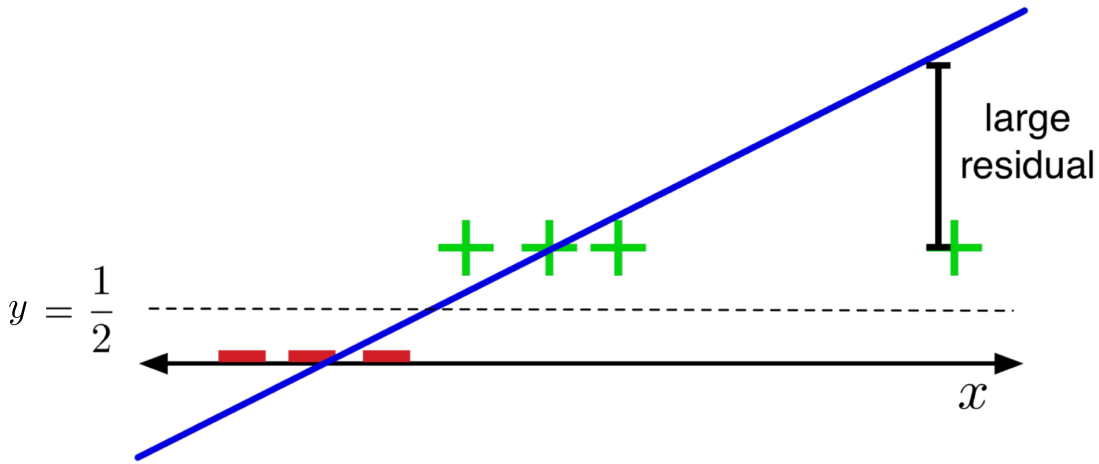



- Q: What is the problem here?

[2]The $\frac{1}{2}$ is just a constant (no effect). Goal: derivative of a pow2 conveniently cancels to 1.

# loss wants to be exact

The loss used was a square[2] loss: $\mathcal{L}_{\text{SE}} = \frac{1}{2}(y-t)^2$ (SE: Squared Error)

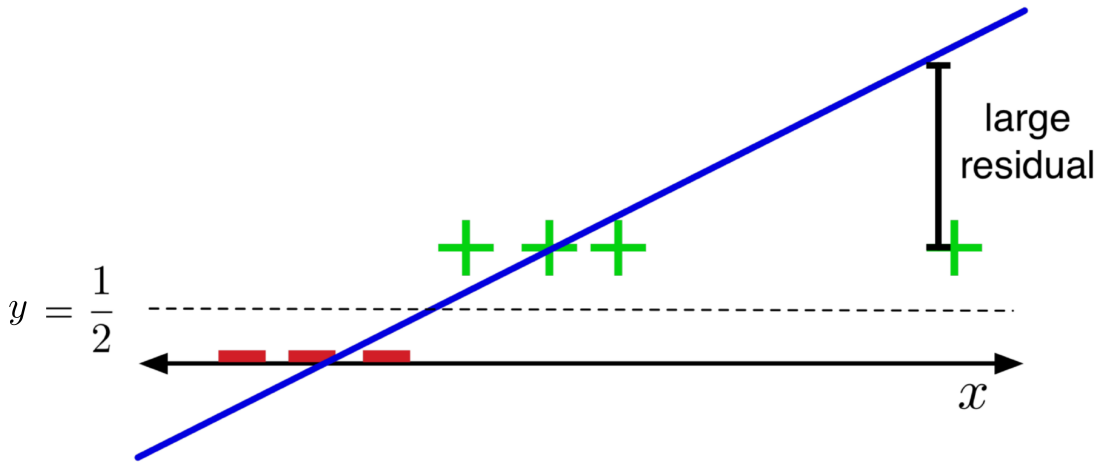



- Q: What is the problem here?
  A: Large losses (residual) for predictions with high confidence.

[2] The $\frac{1}{2}$ is just a constant (no effect). Goal: derivative of a pow2 conveniently cancels to 1.

# loss wants to be exact

The loss used was a square[2] loss: $\mathcal{L}_{\text{SE}} = \frac{1}{2}(y-t)^2$ (SE: Squared Error)




- Q: What is the problem here?
  A: Large losses (residual) for predictions with high confidence.
- If $t = 1$, loss is higher for $y = 10$ then for $y = 0$.

[2]The $\frac{1}{2}$ is just a constant (no effect). Goal: derivative of a pow2 conveniently cancels to 1.

# Logistic regression

Eq (13) in MDL-Loss-RGrosse.pdf

Replace squared loss with Bernoulli cross-entropy loss

- Bernoulli: $y$ (if $t = 1$) and $1 - y$ (if $t = 0$), where $t$=true label

Q: What is the Bernoulli cross entropy (CE) loss $\mathcal{L}_{\text{CE}}$?
(minimize the negative log of the likelihood)

# Logistic regression

Eq (13) in MDL-Loss-RGrosse.pdf

Replace squared loss with Bernoulli cross-entropy loss

- Bernoulli: $y$ (if $t = 1$) and $1 - y$ (if $t = 0$), where $t$=true label

Q: What is the Bernoulli cross entropy (CE) loss $\mathcal{L}_{\text{CE}}$?
(minimize the negative log of the likelihood)

$$\mathcal{L}_{\text{CE}}(y, t) = \begin{cases} -\log y & \text{if } t = 1 \\ -\log(1 - y) & \text{if } t = 0 \end{cases}$$

Same as: $-t \log y - (1 - t) \log(1 - y)$

# Logistic regression

Eq (13) in MDL-Loss-RGrosse.pdf

Replace squared loss with Bernoulli cross-entropy loss

- Bernoulli: $y$ (if $t = 1$) and $1 - y$ (if $t = 0$), where $t$=true label

Q: What is the Bernoulli cross entropy (CE) loss $\mathcal{L}_{\text{CE}}$?
(minimize the negative log of the likelihood)

$$\mathcal{L}_{\text{CE}}(y, t) = \begin{cases} -\log y & \text{if } t = 1 \\ -\log(1 - y) & \text{if } t = 0 \end{cases}$$

Same as: $-t \log y - (1 - t) \log(1 - y)$

Q: Why is this the same?

# Logistic regression

Eq (13) in MDL-Loss-RGrosse.pdf

Replace squared loss with Bernoulli cross-entropy loss

- Bernoulli: $y$ (if $t = 1$) and $1 - y$ (if $t = 0$), where $t$=true label

Q: What is the Bernoulli cross entropy (CE) loss $\mathcal{L}_{\text{CE}}$?
(minimize the negative log of the likelihood)

$$\mathcal{L}_{\text{CE}}(y, t) = \begin{cases} -\log y & \text{if } t = 1 \\ -\log(1 - y) & \text{if } t = 0 \end{cases}$$

Same as: $-t \log y - (1 - t) \log(1 - y)$

Q: Why is this the same?
Q: Draw graphs for $t = 1$ and $t = 0$ ?

# Logistic regression

Eq (13) in MDL-Loss-RGrosse.pdf

Replace squared loss with Bernoulli cross-entropy loss

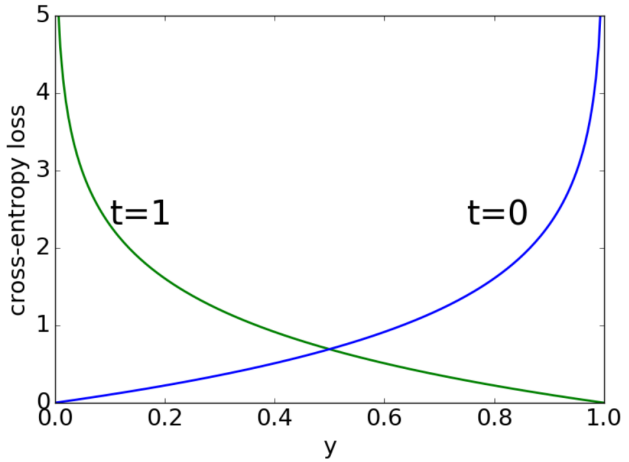
- Bernoulli: $y$ (if $t = 1$) and $1 - y$ (if $t = 0$), where $t$=true label

Q: What is the Bernoulli cross entropy (CE) loss $\mathcal{L}_{\text{CE}}$?
(minimize the negative log of the likelihood)

$$\mathcal{L}_{\text{CE}}(y, t) = \begin{cases} -\log y & \text{if } t = 1 \\ -\log(1 - y) & \text{if } t = 0 \end{cases}$$

Same as: $-t \log y - (1 - t) \log(1 - y)$

Q: Why is this the same?
Q: Draw graphs for $t = 1$ and $t = 0$ ?




$t$=true label

Q: What is penalized most?

# Logistic regression

Eq (13) in MDL-Loss-RGrosse.pdf

Replace squared loss with Bernoulli cross-entropy loss

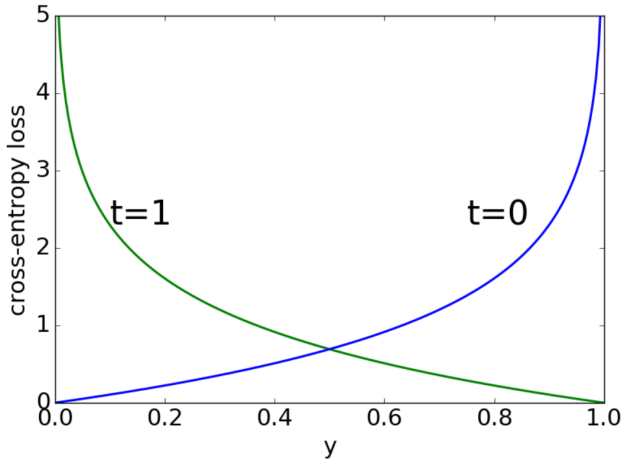
- Bernoulli: $y$ (if $t = 1$) and $1 - y$ (if $t = 0$), where $t$=true label

Q: What is the Bernoulli cross entropy (CE) loss $\mathcal{L}_{\text{CE}}$?
(minimize the negative log of the likelihood)

$$\mathcal{L}_{\text{CE}}(y, t) = \begin{cases} -\log y & \text{if } t = 1 \\ -\log(1 - y) & \text{if } t = 0 \end{cases}$$

Same as: $-t \log y - (1 - t) \log(1 - y)$

Q: Why is this the same?
Q: Draw graphs for $t = 1$ and $t = 0$ ?




$t$=true label

Q: What is penalized most?
A: Penalizes small difference of really wrong predictions

# Questions?

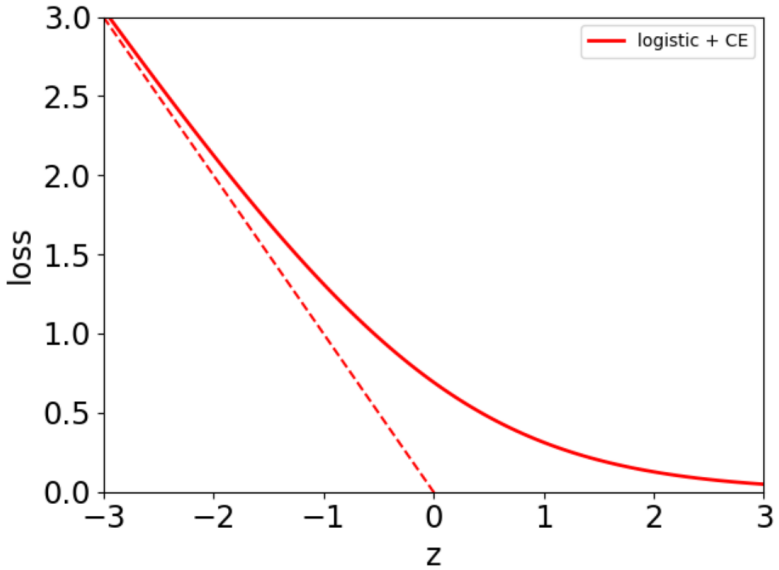
# Logistic regression

Eq (15) in MDL-Loss-RGrosse.pdf

$$z = w^\top x + b$$

$$y = \sigma(z) = \frac{1}{1 + e^{-z}}$$

$$\mathcal{L}_{\text{CE}} = -t \log y - (1 - t) \log(1 - y)$$




Q: Explain these terms

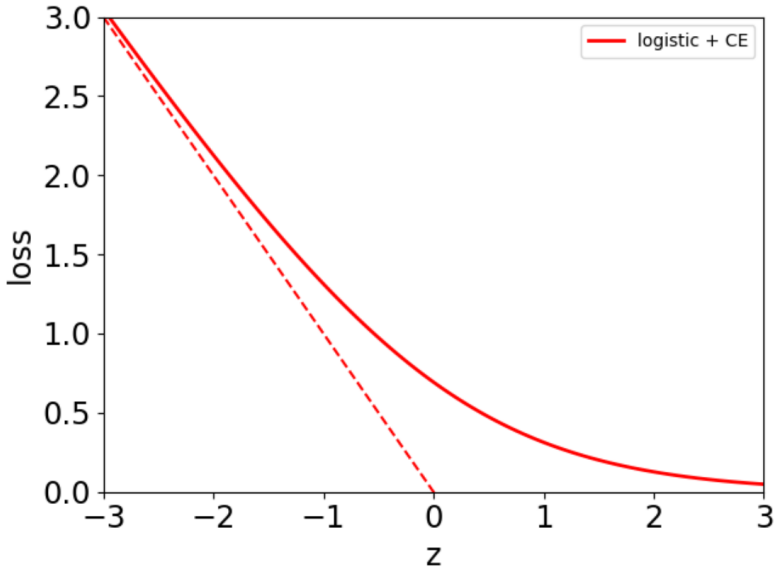
# Logistic regression

Eq (15) in MDL-Loss-RGrosse.pdf

$$z = w^\top x + b$$

$$y = \sigma(z) = \frac{1}{1 + e^{-z}}$$

$$\mathcal{L}_{\text{CE}} = -t \log y - (1 - t) \log(1 - y)$$




Q: Explain these terms

Q: What is shown on this graph?

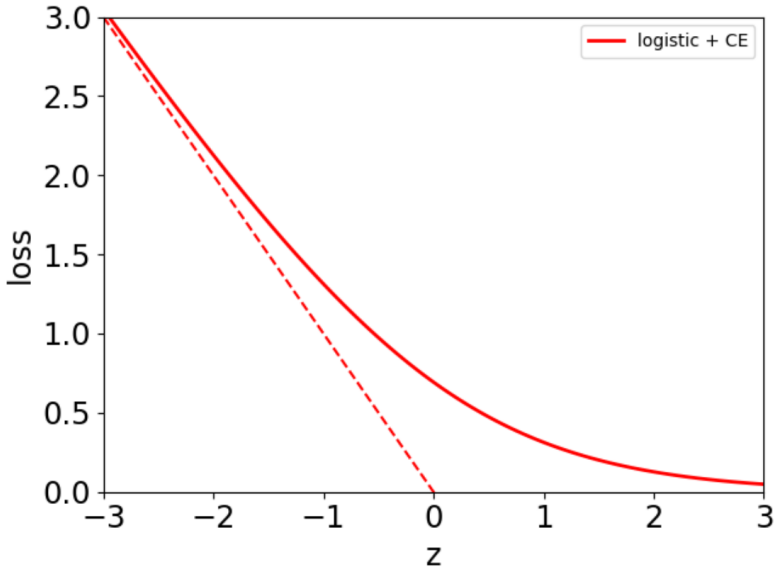
# Logistic regression

Eq (15) in MDL-Loss-RGrosse.pdf

$$z = w^\top x + b$$

$$y = \sigma(z) = \frac{1}{1 + e^{-z}}$$

$$\mathcal{L}_{\text{CE}} = -t \log y - (1 - t) \log(1 - y)$$




Q: Explain these terms

Q: What is shown on this graph?

The loss for $t = 1$ in terms of $z$

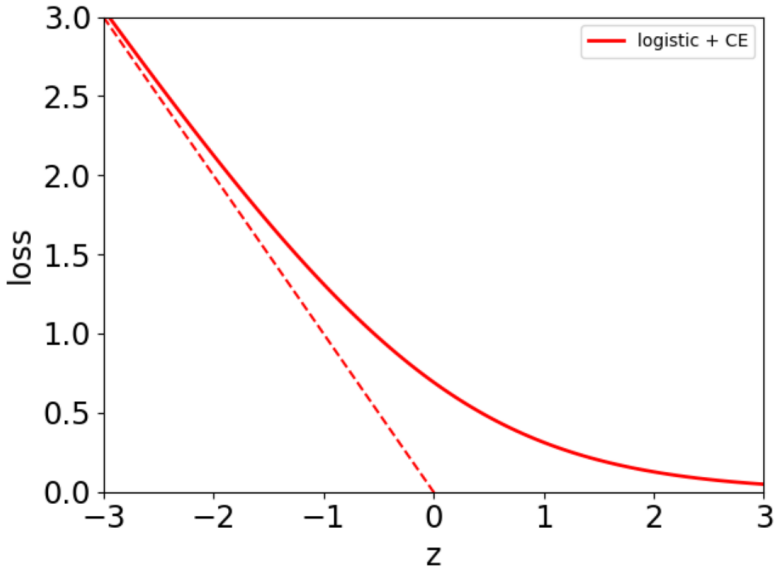
# Logistic regression

Eq (15) in MDL-Loss-RGrosse.pdf

$$z = w^\top x + b$$

$$y = \sigma(z) = \frac{1}{1 + e^{-z}}$$

$$\mathcal{L}_{\text{CE}} = -t \log y - (1 - t) \log(1 - y)$$




Q: Explain these terms

Q: What is shown on this graph?

The loss for $t = 1$ in terms of $z$

Q: What does it penalize?

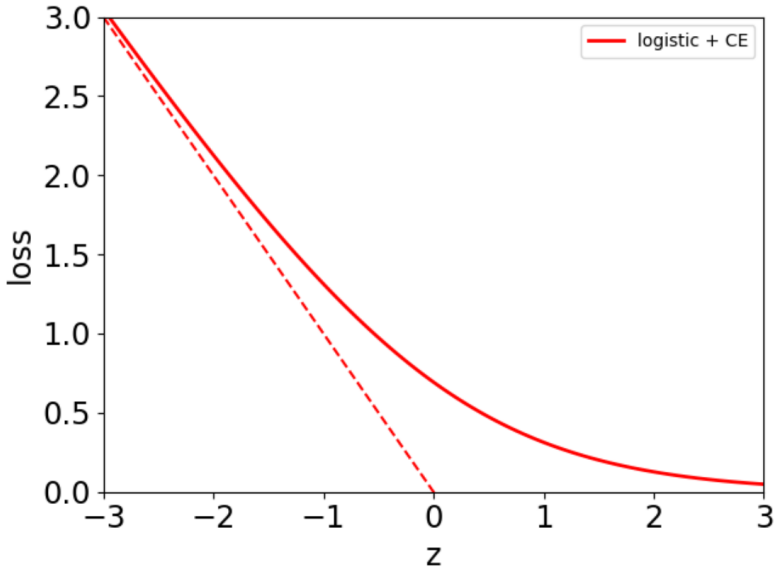
# Logistic regression

Eq (15) in MDL-Loss-RGrosse.pdf

$$z = w^\top x + b$$

$$y = \sigma(z) = \frac{1}{1 + e^{-z}}$$

$$\mathcal{L}_{\text{CE}} = -t \log y - (1 - t) \log(1 - y)$$




Q: Explain these terms

Q: What is shown on this graph?

The loss for $t = 1$ in terms of $z$

Q: What does it penalize?

Penalizes very wrong predictions

# Multiclass classification

DL-book: 6.2.2.3, Eqs (21), (23) in MDL-Loss-RGrosse.pdf; UDL-book: 5.5

- Targets from a discrete set $\{1, \ldots, K\}$
- Convenient: "One-of-K", or "One-hot" encoding:

$$t = \underbrace{(0, \ldots, 0, 1, 0, \ldots, 0)}_{\text{Entry } k \text{ is set to 1}}$$

# Multiclass classification

DL-book: 6.2.2.3, Eqs (21), (23) in MDL-Loss-RGrosse.pdf; UDL-book: 5.5

- Targets from a discrete set $\{1, \ldots, K\}$
- Convenient: "One-of-K", or "One-hot" encoding:

$$t = \underbrace{(0, \ldots, 0, 1, 0, \ldots, 0)}_{\text{Entry } k \text{ is set to 1}}$$

- Softmax: generalization of logistic function:

$$Y_k = \text{softmax}(z_1, z_2, \ldots, z_K) = \frac{e^{z_k}}{\sum_{k'} e^{z_{k'}}}$$

# Multiclass classification

DL-book: 6.2.2.3, Eqs (21), (23) in MDL-Loss-RGrosse.pdf; UDL-book: 5.5

- Targets from a discrete set $\{1, \ldots, K\}$
- Convenient: "One-of-K", or "One-hot" encoding:

$$t = \underbrace{(0, \ldots, 0, 1, 0, \ldots, 0)}_{\text{Entry } k \text{ is set to 1}}$$

- Softmax: generalization of logistic function:

$$Y_k = \text{softmax}(z_1, z_2, \ldots, z_K) = \frac{e^{z_k}}{\sum_{k'} e^{z_{k'}}}$$

- Vector of class probabilities, use cross-entropy as loss function

$$\mathcal{L}_{\text{CE}}(y, t) = -\sum_{k=1}^{K} t_k \log y_k = -t^\top (\log y)$$

- Outputs positive and sum to 1. (Probabilistic interpretation)
- If one of the $z_k$ is much larger, it approximates the arg max.

# Questions?