Qual Presentation

Daniel Khashabi

Outline

- My own line of research
- Papers:
  - Fast Dropout training, ICML, 2013
  - Distributional Semantics Beyond Words: Supervised Learning of Analogy and Paraphrase, TACL, 2013.

Outline

- My own line of research
- Papers:
  - Fast Dropout training, ICML, 2013
  - Distributional Semantics Beyond Words: Supervised Learning of Analogy and Paraphrase, TACL, 2013.

Outline

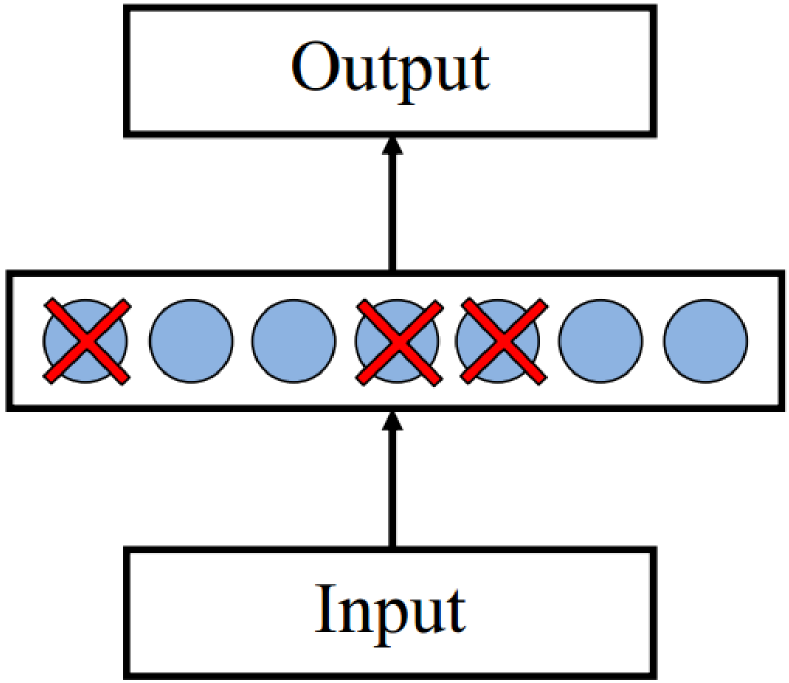
- My own line of research
- Papers:
  - Fast Dropout training, ICML, 2013
  - Distributional Semantics Beyond Words: Supervised Learning of Analogy and Paraphrase, TACL, 2013.

Outline

- My own line of research
- Papers:
  - Fast Dropout training, ICML, 2013
  - Distributional Semantics Beyond Words: Supervised Learning of Analogy and Paraphrase, TACL, 2013.

Dropout training

- Proposed by (Hinton et al, 2012)



- Each time decide whether to delete one hidden unit with some probability p

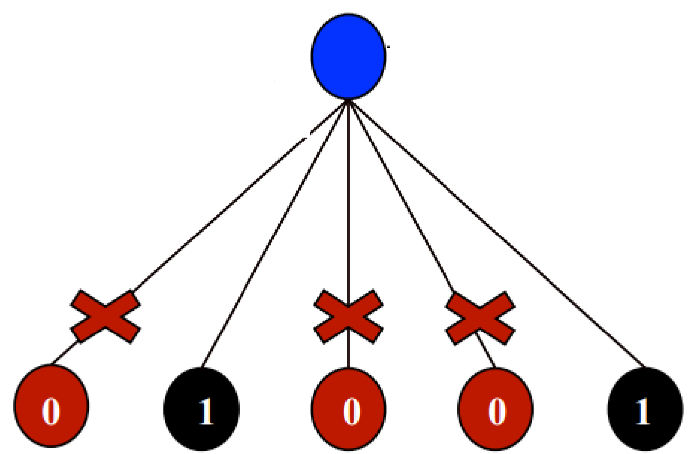
Dropout training

- Model averaging effect
  - Among 2^H models, with shared parameters
  - Only a few get trained
  - Much stronger than the known regularizer

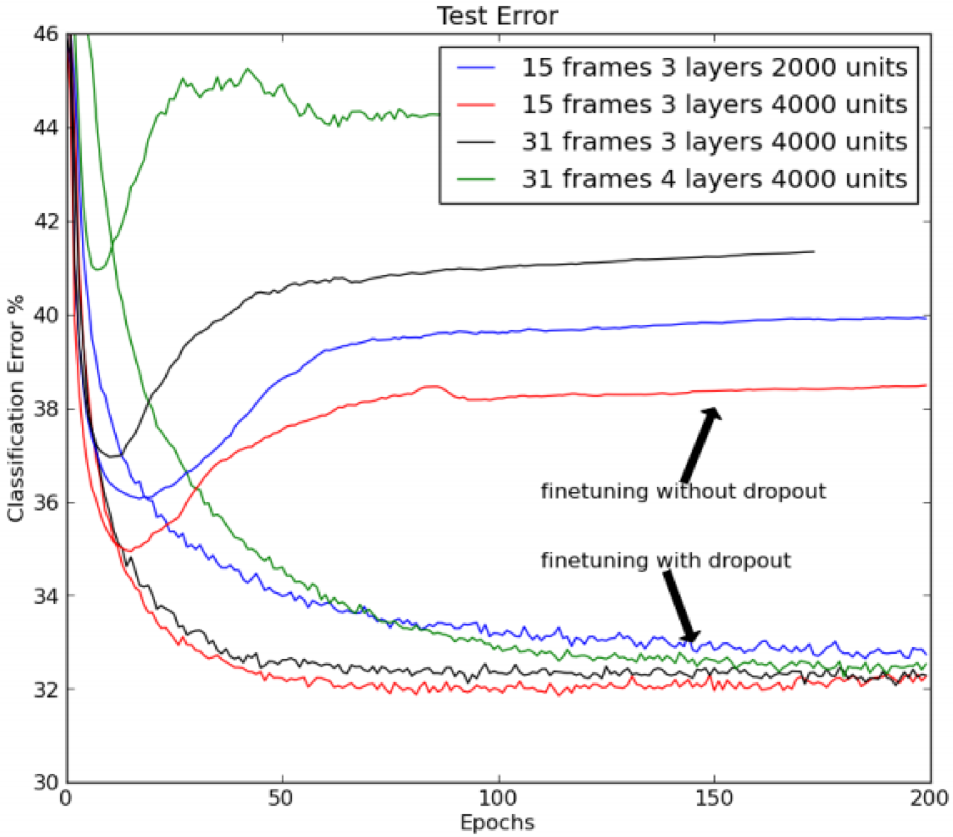
Dropout training

- Model averaging effect
  - Among 2^H models, with shared parameters
  - Only a few get trained
  - Much stronger than the known regularizer

- What about the input space?
  - Do the same thing!



Dropout training



- Dropout of 50% of the hidden units and 20% of the input units (Hinton et al, 2012)

Outline

Outline

- Can we explicitly show that dropout acts as a regularizer?

Outline

- Can we explicitly show that dropout acts as a regularizer?
  - Very easy to show for linear regression

Outline

- Can we explicitly show that dropout acts as a regularizer?
  - Very easy to show for linear regression
  - What about others?

Outline

- Can we explicitly show that dropout acts as a regularizer?
  - Very easy to show for linear regression
  - What about others?
- Dropout needs sampling

Outline

- Can we explicitly show that dropout acts as a regularizer?
  - Very easy to show for linear regression
  - What about others?
- Dropout needs sampling
  - Can be slow!

Outline

- Can we explicitly show that dropout acts as a regularizer?
  - Very easy to show for linear regression
  - What about others?
- Dropout needs sampling
  - Can be slow!
- Can we convert the sampling based update into a deterministic form?
  - ↓

Outline

- Can we explicitly show that dropout acts as a regularizer?
  - Very easy to show for linear regression
  - What about others?
- Dropout needs sampling
  - Can be slow!
- Can we convert the sampling based update into a deterministic form?
  - Find expected form of updates

Linear Regression

Linear Regression

- Reminder:

Linear Regression

- Reminder:

$$z_i \sim Bernoulli(p_i)$$

Linear Regression

- Reminder:

$$z_i \sim Bernoulli(p_i)$$

$$\Rightarrow \mathbf{E}\left[z_i\right] = p_i$$

Linear Regression

- Reminder:

$$z_i \sim Bernoulli(p_i)$$

$$\Rightarrow \mathbf{E}\left[z_i\right] = p_i$$

$$\Rightarrow \mathbf{Var}\left[z_i\right] = p_i(1 - p_i)$$

Linear Regression

- Reminder:

$$z_i \sim Bernoulli(p_i)$$

$$\Rightarrow \mathbf{E}\left[z_i\right] = p_i$$

$$\Rightarrow \mathbf{Var}\left[z_i\right] = p_i(1 - p_i)$$

- Consider the standard linear regression

Linear Regression

- Reminder:

$$z_i \sim Bernoulli(p_i)$$

$$\Rightarrow \mathbf{E}\left[z_i\right] = p_i$$

$$\Rightarrow \mathbf{Var}\left[z_i\right] = p_i(1 - p_i)$$

- Consider the standard linear regression

$$g = w^T x$$

Linear Regression

- Reminder:

$$z_i \sim Bernoulli(p_i)$$

$$\Rightarrow \mathbf{E}\left[z_i\right] = p_i$$

$$\Rightarrow \mathbf{Var}\left[z_i\right] = p_i(1 - p_i)$$

- Consider the standard linear regression

$$g = w^T x$$

$$w^* = \arg\min_w \sum_i \left(w^T x^{(i)} - y^{(i)}\right)^2$$

Linear Regression

- Reminder:

$$z_i \sim Bernoulli(p_i)$$

$$\Rightarrow \mathbf{E}\left[z_i\right] = p_i$$

$$\Rightarrow \mathbf{Var}\left[z_i\right] = p_i(1 - p_i)$$

- Consider the standard linear regression

$$g = w^T x$$

$$w^* = \arg\min_w \sum_i \left(w^T x^{(i)} - y^{(i)}\right)^2$$

- With regularization:

$$L(w) = \sum_i \left(w^T x^{(i)} - y^{(i)}\right)^2 + \lambda \sum_i w_i^2$$

Linear Regression

- Reminder:

$$z_i \sim Bernoulli(p_i)$$

$$\Rightarrow \mathbf{E}\left[z_i\right] = p_i$$

$$\Rightarrow \mathbf{Var}\left[z_i\right] = p_i(1 - p_i)$$

- Consider the standard linear regression

$$g = w^T x$$

$$w^* = \arg\min_w \sum_i \left(w^T x^{(i)} - y^{(i)}\right)^2$$

- With regularization:

$$L(w) = \sum_i \left(w^T x^{(i)} - y^{(i)}\right)^2 + \lambda \sum_i w_i^2$$

- Closed form solution:

$$w = \left(X^T X + \lambda I\right)^{-1} X^T y$$

Dropout Linear Regression

- Consider the standard linear regression

$$g = w^T x$$

Dropout Linear Regression

- Consider the standard linear regression

$$g = w^T x$$

$$x_i \Leftrightarrow z_i \sim Bernoulli(p_i)$$

Dropout Linear Regression

- Consider the standard linear regression

$$g = w^T x$$

$$x_i \Leftrightarrow z_i \sim Bernoulli(p_i)$$

$$D_z = diag(z_1, \ldots, z_m)$$

Dropout Linear Regression

- Consider the standard linear regression

$$g = w^T x$$

$$x_i \Leftrightarrow z_i \sim Bernoulli(p_i)$$

$$D_z = diag(z_1, ..., z_m)$$

- LR with dropout:

$$w^T D_z x$$

Dropout Linear Regression

- Consider the standard linear regression

$$g = w^T x$$

$$x_i \Leftrightarrow z_i \sim Bernoulli(p_i)$$

$$D_z = diag(z_1, \ldots, z_m)$$

- LR with dropout:

$$w^T \boxed{D_z} x$$

Dropout Linear Regression

- Consider the standard linear regression

$$g = w^T x$$

$$x_i \Leftrightarrow z_i \sim Bernoulli(p_i)$$

$$D_z = diag(z_1, \ldots, z_m)$$

- LR with dropout:

$$w^T \boxed{D_z} x$$

- How to find the parameter?

$$L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)} \right)^2$$

Dropout Linear Regression

- Consider the standard linear regression

$$g = w^T x$$

$$x_i \Leftrightarrow z_i \sim Bernoulli(p_i)$$

$$D_z = diag(z_1, \ldots, z_m)$$

- LR with dropout:

$$w^T D_z x$$

- How to find the parameter?

$$L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)} \right)^2$$

Fast Dropout for Linear Regression

- We had: $$L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)} \right)^2$$
- Instead of sampling, minimize the expected loss
  - Fixed *x* and *y*: $$\mathbf{E}\left[\left(w^T D_z x - y \right)^2 \right]$$
-

Fast Dropout for Linear Regression

- We had: $$L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)} \right)^2$$
- Instead of sampling, minimize the expected loss
  - Fixed x and y: $$\mathbf{E}\left[\left(w^T D_z x - y \right)^2 \right]$$
-

Fast Dropout for Linear Regression

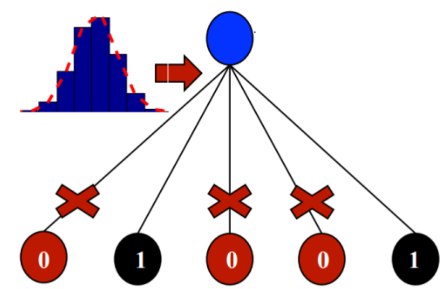
- We had: $L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)} \right)^2$
- Instead of sampling, minimize the expected loss
  - Fixed x and y: $\mathbf{E}\left[\left(w^T D_z x - y \right)^2 \right]$
-

Fast Dropout for Linear Regression

- We had: $L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)} \right)^2$
- Instead of sampling, minimize the expected loss
  - Fixed x and y: $\mathbf{E}\left[\left(w^T D_z x - y \right)^2 \right]$
-

Fast Dropout for Linear Regression

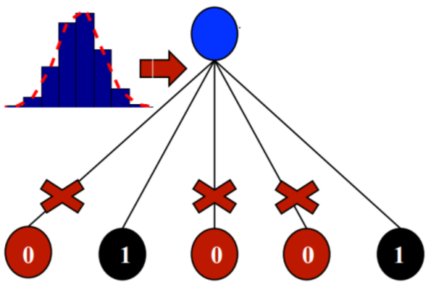
- We had: $$L(w) = \sum_i \left(w^T \boxed{D_z} x^{(i)} - y^{(i)} \right)^2$$
- Instead of sampling, minimize the expected loss
  - Fixed x and y: $$\mathbf{E}\left[\left(\boxed{w^T D_z x} - y \right)^2 \right]$$



- $$w^T D_z x = \sum_{i=1}^{m} w_i x_i z_i \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$$

Fast Dropout for Linear Regression

- We had: $$L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)} \right)^2$$
- Instead of sampling, minimize the expected loss
  - Fixed x and y: $$\mathbf{E}\left[\left(w^T D_z x - y \right)^2 \right]$$

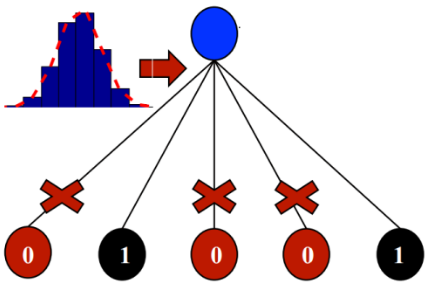


- $$w^T D_z x = \sum_{i=1}^{m} w_i x_i z_i \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$$

$$\mu_S =$$

Fast Dropout for Linear Regression

- We had: $$L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)} \right)^2$$
- Instead of sampling, minimize the expected loss
  - Fixed x and y: $$\mathbf{E}\left[\left(w^T D_z x - y \right)^2 \right]$$

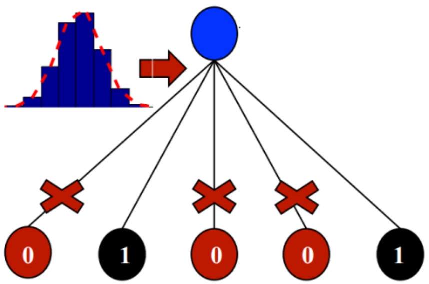


- $$w^T D_z x = \sum_{i=1}^{m} w_i x_i z_i \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$$

  $$\mu_S =$$

Fast Dropout for Linear Regression

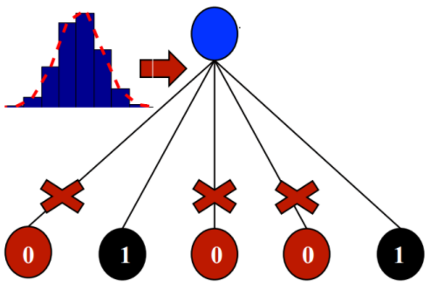
- We had: $$L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)} \right)^2$$
- Instead of sampling, minimize the expected loss
  - Fixed x and y: $$\mathbf{E}\left[\left(w^T D_z x - y \right)^2 \right]$$



- $$w^T D_z x = \sum_{i=1}^{m} w_i x_i z_i \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$$
  $$\mu_S = \mathbf{E}\left[w^T D_z x \right]$$

Fast Dropout for Linear Regression

- We had: $$L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)} \right)^2$$
- Instead of sampling, minimize the expected loss
  - Fixed x and y: $$\mathbf{E}\left[\left(w^T D_z x - y \right)^2 \right]$$

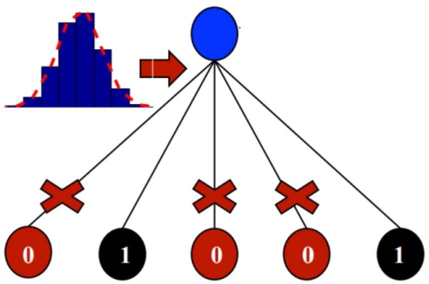


- $$w^T D_z x = \sum_{i=1}^{m} w_i x_i z_i \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$$

$$\mu_S = \mathbf{E}\left[w^T D_z x \right] = \sum_{i=1}^{m} w_i x_i \mathbf{E}[z_i]$$

Fast Dropout for Linear Regression

- We had: $$L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)} \right)^2$$
- Instead of sampling, minimize the expected loss
  - Fixed x and y: $$\mathbf{E}\left[\left(w^T D_z x - y \right)^2 \right]$$

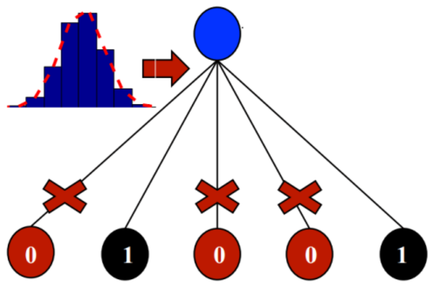


- $$w^T D_z x = \sum_{i=1}^{m} w_i x_i z_i \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$$

$$\mu_S = \mathbf{E}\left[w^T D_z x \right] = \sum_{i=1}^{m} w_i x_i \mathbf{E}[z_i] = \sum_{i=1}^{m} w_i x_i p_i$$

Fast Dropout for Linear Regression

- We had: $$L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)} \right)^2$$
- Instead of sampling, minimize the expected loss
  - Fixed x and y: $$\mathbf{E}\left[\left(w^T D_z x - y \right)^2 \right]$$



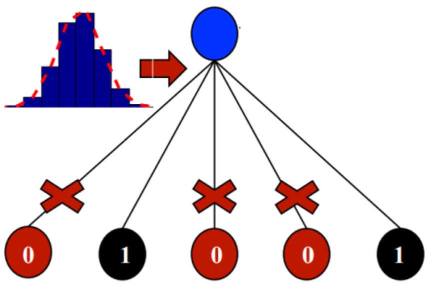
- $$w^T D_z x = \sum_{i=1}^{m} w_i x_i z_i \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$$

$$\mu_S = \mathbf{E}\left[w^T D_z x \right] = \sum_{i=1}^{m} w_i x_i \mathbf{E}[z_i] = \sum_{i=1}^{m} w_i x_i p_i$$

$$\sigma_S^2 =$$

Fast Dropout for Linear Regression

- We had: $$L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)}\right)^2$$
- Instead of sampling, minimize the expected loss
  - Fixed x and y: $$\mathbf{E}\left[\left(w^T D_z x - y\right)^2\right]$$



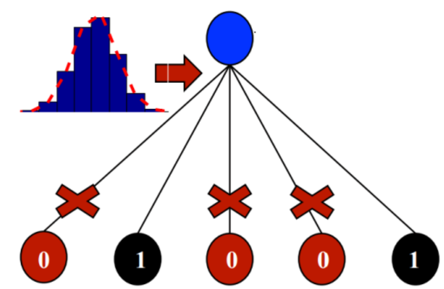
- $$w^T D_z x = \sum_{i=1}^{m} w_i x_i z_i \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$$

$$\mu_S = \mathbf{E}\left[w^T D_z x\right] = \sum_{i=1}^{m} w_i x_i \mathbf{E}[z_i] = \sum_{i=1}^{m} w_i x_i p_i$$

$$\sigma_S^2 =$$

Fast Dropout for Linear Regression

- We had: $L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)} \right)^2$
- Instead of sampling, minimize the expected loss
  - Fixed x and y: $\mathbf{E}\left[\left(w^T D_z x - y \right)^2 \right]$



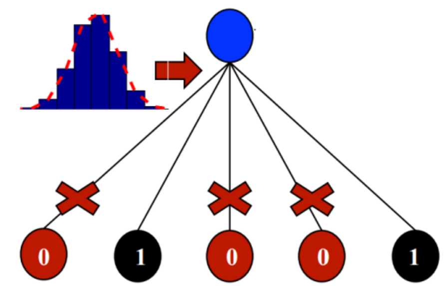
- $$w^T D_z x = \sum_{i=1}^{m} w_i x_i z_i \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$$

$$\mu_S = \mathbf{E}\left[w^T D_z x \right] = \sum_{i=1}^{m} w_i x_i \mathbf{E}\left[z_i \right] = \sum_{i=1}^{m} w_i x_i p_i$$

$$\sigma_S^2 = \mathbf{Var}\left[w^T D_z x \right] =$$

Fast Dropout for Linear Regression

- We had: $L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)} \right)^2$
- Instead of sampling, minimize the expected loss
  - Fixed x and y: $\mathbf{E}\left[\left(w^T D_z x - y \right)^2 \right]$



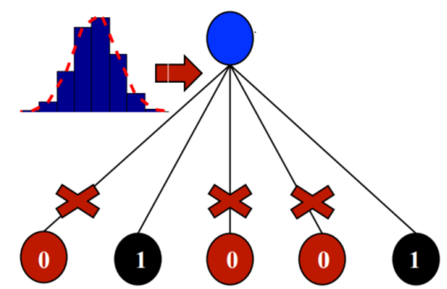
- $$w^T D_z x = \sum_{i=1}^{m} w_i x_i z_i \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$$

$$\mu_S = \mathbf{E}\left[w^T D_z x \right] = \sum_{i=1}^{m} w_i x_i \mathbf{E}[z_i] = \sum_{i=1}^{m} w_i x_i p_i$$

$$\sigma_S^2 = \mathbf{Var}\left[w^T D_z x \right] = \sum_{i=1}^{m} (w_i x_i)^2 \mathbf{Var}[z_i] =$$

Fast Dropout for Linear Regression

- We had: $L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)} \right)^2$
- Instead of sampling, minimize the expected loss
  - Fixed x and y: $\mathbf{E}\left[\left(w^T D_z x - y \right)^2 \right]$



- $$w^T D_z x = \sum_{i=1}^{m} w_i x_i z_i \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$$

$$\mu_S = \mathbf{E}\left[w^T D_z x \right] = \sum_{i=1}^{m} w_i x_i \mathbf{E}[z_i] = \sum_{i=1}^{m} w_i x_i p_i$$

$$\sigma_S^2 = \mathbf{Var}\left[w^T D_z x \right] = \sum_{i=1}^{m} (w_i x_i)^2 \, \mathbf{Var}[z_i] = \sum_{i=1}^{m} (w_i x_i)^2 \, p_i (1 - p_i)$$

Fast Dropout for Linear Regression

- We had: $$L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)} \right)^2$$
- Instead of sampling minimize the expected loss:

$$w^T D_z x = \sum_{i=1}^{m} w_i x_i z_i \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$$

Fast Dropout for Linear Regression

- We had: $L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)} \right)^2$
- Instead of sampling minimize the expected loss:

$$w^T D_z x = \sum_{i=1}^{m} w_i x_i z_i \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$$

$$\mathbf{E}\left[\left(w^T D_z x - y \right)^2 \right] \simeq \mathbf{E}_{S \sim N(\mu_S, \sigma_S^2)}\left[\left(S - y \right)^2 \right]$$

↓

Fast Dropout for Linear Regression

- We had: $$L(w)=\sum_i\left(w^T D_z x^{(i)} - y^{(i)}\right)^2$$
- Instead of sampling minimize the expected loss:

$$w^T D_z x = \sum_{i=1}^{m} w_i x_i z_i \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$$

$$\mathbf{E}\left[\left(w^T D_z x - y\right)^2\right] \simeq \mathbf{E}_{S\sim N(\mu_S,\sigma_S^2)}\left[(S-y)^2\right] = \left(\mu_S - y\right)^2 + \sigma_S^2$$

-

Fast Dropout for Linear Regression

- We had: $L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)} \right)^2$
- Instead of sampling minimize the expected loss:

$$w^T D_z x = \sum_{i=1}^{m} w_i x_i z_i \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$$

$$\mathbf{E}\left[\left(w^T D_z x - y \right)^2 \right] \simeq \mathbf{E}_{S \sim N(\mu_S, \sigma_S^2)} \left[(S - y)^2 \right] = \left(\mu_S - y \right)^2 + \sigma_S^2$$

Fast Dropout for Linear Regression

- We had: $L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)} \right)^2$
- Instead of sampling minimize the expected loss:

$$w^T D_z x = \sum_{i=1}^{m} w_i x_i z_i \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$$

$$\mathbf{E}\left[\left(w^T D_z x - y \right)^2 \right] \simeq \mathbf{E}_{S \sim N(\mu_S, \sigma_S^2)} \left[(S - y)^2 \right] = \left(\mu_S - y \right)^2 + \sigma_S^2$$

Fast Dropout for Linear Regression

- We had: $$L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)} \right)^2$$
- Instead of sampling minimize the expected loss:

$$w^T D_z x = \sum_{i=1}^{m} w_i x_i z_i \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$$

$$\mathbf{E}\left[\left(w^T D_z x - y \right)^2 \right] \simeq \mathbf{E}_{S \sim N(\mu_S, \sigma_S^2)}\left[(S - y)^2 \right] = \left(\mu_S - y \right)^2 + \sigma_S^2$$

- Expected loss:

Fast Dropout for Linear Regression

- We had: $L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)} \right)^2$
- Instead of sampling minimize the expected loss:

$$w^T D_z x = \sum_{i=1}^{m} w_i x_i z_i \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$$

$$\mathbf{E}\left[\left(w^T D_z x - y \right)^2 \right] \simeq \mathbf{E}_{S \sim N(\mu_S, \sigma_S^2)} \left[\left(S - y \right)^2 \right] = \left(\mu_S - y \right)^2 + \sigma_S^2$$

- Expected loss:

$$\tilde{L}(w) = \mathbf{E} L(w) \simeq \sum_i \left(\mu_S^{(i)} - y^{(i)} \right)^2 + \sigma_S^{2\,(i)}$$

Fast Dropout for Linear Regression

- We had: $$L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)} \right)^2$$
- Instead of sampling minimize the expected loss:

$$w^T D_z x = \sum_{i=1}^{m} w_i x_i z_i \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$$

$$\mathbf{E}\left[\left(w^T D_z x - y \right)^2 \right] \simeq \mathbf{E}_{S \sim N(\mu_S, \sigma_S^2)} \left[(S - y)^2 \right] = \left(\mu_S - y \right)^2 + \sigma_S^2$$

- Expected loss:

$$\tilde{L}(w) = \mathbf{E} L(w) \simeq \sum_i \left(\mu_S^{(i)} - y^{(i)} \right)^2 + \sigma_S^{2\,(i)}$$

$$= \sum_i \left(\mu_S^{(i)} - y^{(i)} \right)^2 + \lambda \sum_i c_i w_i^2, \quad c_i = \sum_j \left(x_i^{(j)} \right)^2$$

Fast Dropout for Linear Regression

- We had: $L(w) = \sum_i \left(w^T D_z x^{(i)} - y^{(i)} \right)^2$
- Instead of sampling minimize the expected loss:

$$w^T D_z x = \sum_{i=1}^{m} w_i x_i z_i \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$$

$$\mathbf{E}\left[\left(w^T D_z x - y \right)^2 \right] \simeq \mathbf{E}_{S \sim N(\mu_S, \sigma_S^2)} \left[\left(S - y \right)^2 \right] = \left(\mu_S - y \right)^2 + \sigma_S^2$$

- Expected loss:

$$\tilde{L}(w) = \mathbf{E} L(w) \simeq \sum_i \left(\mu_S^{(i)} - y^{(i)} \right)^2 + \sigma_S^{2\,(i)}$$

$$= \sum_i \left(\mu_S^{(i)} - y^{(i)} \right)^2 + \lambda \sum_i c_i w_i^2, \quad c_i = \sum_j \left(x_i^{(j)} \right)^2$$

Fast Dropout for Linear Regression

- Expected loss:

$$\tilde{L}(w) = \mathbf{E}L(w) \simeq \sum_i \left(\mu_S^{(i)} - y^{(i)}\right)^2 + \sigma_S^{2\,(i)}$$

$$= \sum_i \left(\mu_S^{(i)} - y^{(i)}\right)^2 + \lambda \sum_i c_i w_i^2, \quad c_i = \sum_j \left(x_i^{(j)}\right)^2$$

- Data-dependent regulizer

Fast Dropout for Linear Regression

- Expected loss:

$$\tilde{L}(w) = \mathbf{E}L(w) \simeq \sum_i \left(\mu_S^{(i)} - y^{(i)}\right)^2 + \sigma_S^{2\,(i)}$$

$$= \sum_i \left(\mu_S^{(i)} - y^{(i)}\right)^2 + \lambda \sum_i c_i w_i^2, \quad c_i = \sum_j \left(x_i^{(j)}\right)^2$$

- Data-dependent regulizer
- Closed form could be found:

$$w = \left(X^T X + \lambda diag(X^T X)\right)^{-1} X^T y$$

Fast Dropout for Linear Regression

- Expected loss:

$$\tilde{L}(w) = \mathbf{E}L(w) \simeq \sum_i \left(\mu_S^{(i)} - y^{(i)}\right)^2 + \sigma_S^{2\,(i)}$$

$$= \sum_i \left(\mu_S^{(i)} - y^{(i)}\right)^2 + \lambda \sum_i c_i w_i^2, \quad c_i = \sum_j \left(x_i^{(j)}\right)^2$$

- Data-dependent regulizer
- Closed form could be found:

$$w = \left(X^T X + \lambda \, diag(X^T X)\right)^{-1} X^T y$$

Some definitions

- Dropout each input dimension randomly:

Some definitions

- Dropout each input dimension randomly:
- Probit:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$

Some definitions

- Dropout each input dimension randomly:
- Probit:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$

- Logistic function / sigmoid :

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

Some definitions useful equalities

-

-

Some definitions useful equalities

- Useful equalities

↓

Some definitions useful equalities

- Useful equalities

$$\int_{-\infty}^{+\infty} \Phi(\lambda x) N(x; \mu, s^2) dx = \Phi\left(\frac{\mu}{\sqrt{s^2 + \lambda^{-2}}}\right)$$

Some definitions useful equalities

- Useful equalities

$$\int_{-\infty}^{+\infty} \Phi(\lambda x) N(x; \mu, s^2) dx = \Phi\left(\frac{\mu}{\sqrt{s^2 + \lambda^{-2}}}\right)$$

$$\sigma(x) \simeq \Phi\left(\sqrt{\frac{\pi}{8}} x\right)$$

Some definitions useful equalities

- Useful equalities

$$\int_{-\infty}^{+\infty} \Phi(\lambda x) N(x; \mu, s^2) dx = \Phi\left(\frac{\mu}{\sqrt{s^2 + \lambda^{-2}}}\right)$$

$$\sigma(x) \simeq \Phi\left(\sqrt{\frac{\pi}{8}} x\right)$$

$$\int_{-\infty}^{+\infty} \sigma(x) N(x; \mu, s^2) dx \simeq \sigma\left(\frac{\mu}{\sqrt{\pi s^2 / 8 + 1}}\right)$$

Some definitions useful equalities

- Useful equalities

$$\int_{-\infty}^{+\infty} \Phi(\lambda x) N(x; \mu, s^2) dx = \Phi\left(\frac{\mu}{\sqrt{s^2 + \lambda^{-2}}}\right)$$

$$\sigma(x) \simeq \Phi\left(\sqrt{\frac{\pi}{8}} x\right)$$

$$\int_{-\infty}^{+\infty} \sigma(x) N(x; \mu, s^2) dx \simeq \sigma\left(\frac{\mu}{\sqrt{\pi s^2 / 8 + 1}}\right)$$

- We can find the following expectation in closed form:

$$\mathbf{E}_{S \sim N(\mu, \sigma^2)} \left[\sigma(S) \right]$$

Logistic Regression

Logistic Regression

- Consider the standard LR

$$P(Y=1 \mid X=x)=\sigma(w^T x)=\frac{1}{1+e^{-w^T x}}$$

Logistic Regression

- Consider the standard LR

$$P(Y=1 \mid X=x) = \sigma(w^T x) = \frac{1}{1+e^{-w^T x}}$$

- The standard gradient update rule is

$$\Delta w_j = (y - \sigma(w^T x))x_j$$

Logistic Regression

- Consider the standard LR

$$P(Y=1 \mid X=x)=\sigma(w^T x)=\frac{1}{1+e^{-w^T x}}$$

- The standard gradient update rule is

$$\Delta w_j=(y-\sigma(w^T x))x_j$$

- For the parameter vector

$$\Delta w_{\log}=(y-\sigma(w^T x))x$$

Dropout on a Logistic Regression

Dropout on a Logistic Regression

- Dropout each input dimension randomly:

↓

↓

Dropout on a Logistic Regression

- Dropout each input dimension randomly:

$$x_i \Leftrightarrow z_i \sim Bernoulli(p_i) \qquad D_z = diag(z_1, \ldots, z_m)$$

Dropout on a Logistic Regression

- Dropout each input dimension randomly:

$$x_i \Leftrightarrow z_i \sim Bernoulli(p_i) \qquad D_z = diag(z_1, \ldots, z_m)$$

- For the parameter vector

$$\Delta w_{\log} = (y - \sigma(w^T x))x \Rightarrow \Delta w = (y - \sigma(w^T D_z x))D_z x$$

Dropout on a Logistic Regression

- Dropout each input dimension randomly:

$$x_i \Leftrightarrow z_i \sim Bernoulli(p_i) \qquad D_z = diag(z_1, \ldots, z_m)$$

- For the parameter vector

$$\Delta w_{\log} = (y - \sigma(w^T x))x \Rightarrow \Delta w = (y - \sigma(w^T D_z x))D_z x$$

Dropout on a Logistic Regression

- Dropout each input dimension randomly:

$$x_i \Leftrightarrow z_i \sim Bernoulli(p_i) \qquad D_z = diag(z_1, \ldots, z_m)$$

- For the parameter vector

$$\Delta w_{\log} = (y - \sigma(w^T x))x \Rightarrow \Delta w = (y - \sigma(w^T D_z x))D_z x$$

- Notation: $x_i = i$-th dimension of x

Dropout on a Logistic Regression

- Dropout each input dimension randomly:

$$x_i \Leftrightarrow z_i \sim Bernoulli(p_i) \qquad D_z = diag(z_1, \ldots, z_m)$$

- For the parameter vector

$$\Delta w_{\log} = (y - \sigma(w^T x))x \Rightarrow \Delta w = (y - \sigma(w^T D_z x))D_z x$$

- Notation:

$$x_i = i\text{-th dimension of } x$$
$$x^{(j)} = j\text{-th training instance}$$

Dropout on a Logistic Regression

- Dropout each input dimension randomly:

$$x_i \Leftrightarrow z_i \sim Bernoulli(p_i) \qquad D_z = diag(z_1, \ldots, z_m)$$

- For the parameter vector

$$\Delta w_{\log} = (y - \sigma(w^T x))x \Rightarrow \Delta w = (y - \sigma(w^T D_z x))D_z x$$

- Notation:

$$x_i = i\text{-th dimension of } x$$
$$x^{(j)} = j\text{-th training instance}$$
$$x_i^{(j)} = i\text{-th dimension of } j\text{-th instance}$$

Dropout on a Logistic Regression

- Dropout each input dimension randomly:

$$x_i \Leftrightarrow z_i \sim Bernoulli(p_i) \qquad D_z = diag(z_1, \ldots, z_m)$$

- For the parameter vector

$$\Delta w_{\log} = (y - \sigma(w^T x))x \Rightarrow \Delta w = (y - \sigma(w^T D_z x))D_z x$$

- Notation:

$$x_i = i\text{-th dimension of } x$$
$$x^{(j)} = j\text{-th training instance}$$
$$x_i^{(j)} = i\text{-th dimension of } j\text{-th instance}$$
$$1 \le i \le m \qquad 1 \le j \le n$$

Fast Dropout training

- Instead of using Δw we use its expectation:

$$\Delta w_{avg} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)} \Delta w$$

Fast Dropout training

- Instead of using Δw we use its expectation:

$$\begin{aligned}\Delta w_{avg} &= \mathbf{E}_{z;z_i \sim Bernoulli(p_i)} \Delta w \\ &= \mathbf{E}_{z;z_i \sim Bernoulli(p_i)} (y - \sigma(w^T D_z x)) D_z x\end{aligned}$$

Fast Dropout training

- Instead of using Δw we use its expectation:

$$\Delta w_{avg} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)} \Delta w$$

$$= \mathbf{E}_{z;z_i \sim Bernoulli(p_i)} (y - \sigma(w^T D_z x)) D_z x$$

Fast Dropout training

- Instead of using Δw we use its expectation:

$$\Delta w_{avg} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)} \Delta w$$
$$= \mathbf{E}_{z;z_i \sim Bernoulli(p_i)} (y - \sigma(w^T D_z x)) D_z x$$

$$w^T D_z x = \sum_{i=1}^{m} w_i x_i z_i \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$$

Fast Dropout training

- Instead of using Δw we use its expectation:

$$\Delta w_{avg} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)} \Delta w$$
$$= \mathbf{E}_{z;z_i \sim Bernoulli(p_i)} (y - \sigma(w^T D_z x)) D_z x$$

$$w^T D_z x = \sum_{i=1}^{m} w_i x_i z_i \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$$

$$\mu_S = \mathbf{E}\left[w^T D_z x\right] = \sum_{i=1}^{m} w_i x_i \mathbf{E}[z_i] = \sum_{i=1}^{m} w_i x_i p_i$$

Fast Dropout training

- Instead of using Δw we use its expectation:

$$\Delta w_{avg} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)} \Delta w$$

$$= \mathbf{E}_{z;z_i \sim Bernoulli(p_i)} (y - \sigma(w^T D_z x)) D_z x$$

$$w^T D_z x = \sum_{i=1}^{m} w_i x_i z_i \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$$

$$\mu_S = \mathbf{E}\left[w^T D_z x\right] = \sum_{i=1}^{m} w_i x_i \mathbf{E}\left[z_i\right] = \sum_{i=1}^{m} w_i x_i p_i$$

$$\sigma_S^2 = \mathbf{Var}\left[w^T D_z x\right] = \sum_{i=1}^{m} (w_i x_i)^2 \,\mathbf{Var}\left[z_i\right] = \sum_{i=1}^{m} (w_i x_i)^2 \, p_i (1 - p_i)$$

Fast Dropout training

- Approx: $\Delta w_{avg} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)}\left[(y-\sigma(w^T D_z x))D_z x\right]$

Fast Dropout training

- Approx: $\Delta w_{avg} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)}\left[(y-\sigma(w^T D_z x))D_z x\right]$
- By knowing: $w^T D_z x \simeq S, \quad S \sim N(\mu_S, \sigma_S^2)$

Fast Dropout training

- Approx: $\Delta w_{avg} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)}\left[(y - \sigma(w^T D_z x))D_z x\right]$
- By knowing: $w^T D_z x \simeq S, \quad S \sim N(\mu_S, \sigma_S^2)$
- How to approximate?

-
-
-

Fast Dropout training

- Approx: $\Delta w_{avg} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)}\left[(y-\sigma(w^T D_z x))D_z x\right]$
- By knowing: $w^T D_z x \simeq S, \quad S \sim N(\mu_S, \sigma_S^2)$
- How to approximate?

- Option 1: $\mathbf{E}_S\left[(y-\sigma(S))\right]\mathbf{E}_z\left[D_z x\right]$

Fast Dropout training

- Approx: $\Delta w_{avg} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)}\left[(y-\sigma(w^T D_z x))D_z x\right]$
- By knowing: $w^T D_z x \simeq S, \quad S \sim N(\mu_S, \sigma_S^2)$
- How to approximate?

- Option 1: $\mathbf{E}_S\left[(y-\sigma(S))\right]\mathbf{E}_z\left[D_z x\right]$
- Option 2: $(y-\sigma(\mathbf{E}_S[S]))\mathbf{E}_z\left[D_z x\right]$

↓

Fast Dropout training

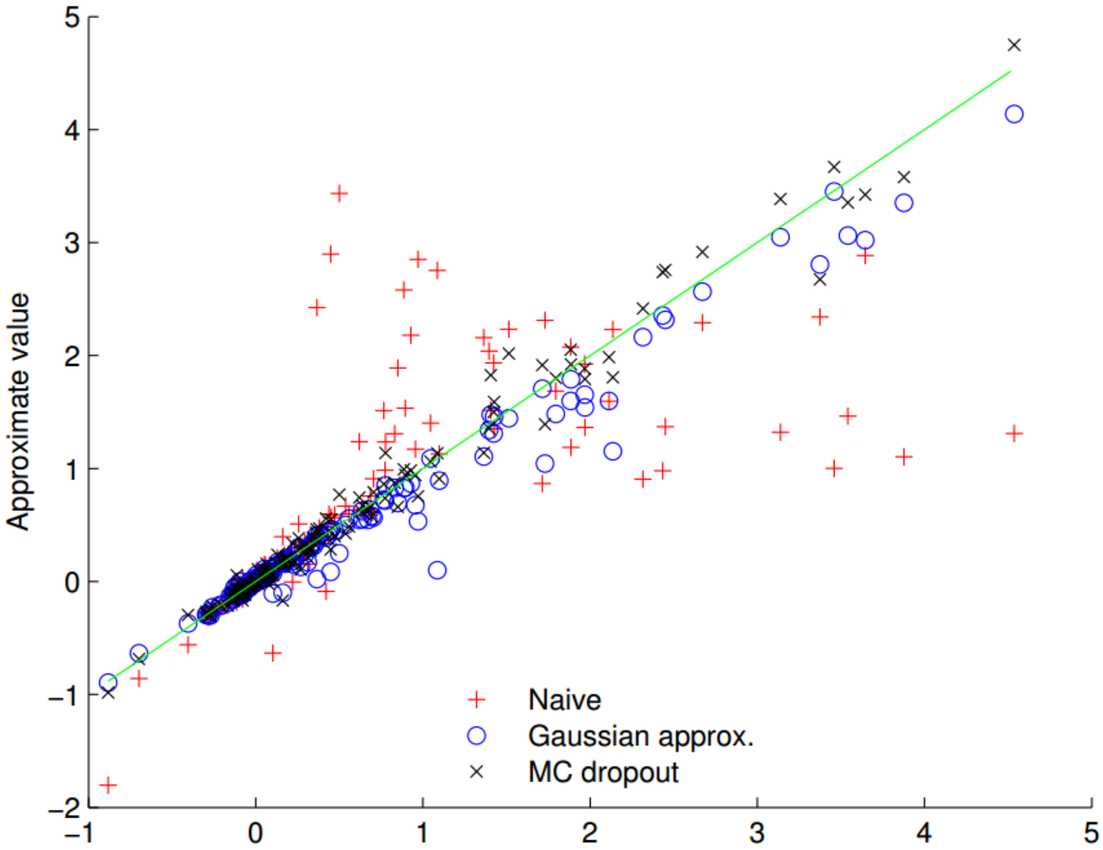
- Approx: $\Delta w_{avg} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)}\left[(y-\sigma(w^T D_z x))D_z x\right]$
- By knowing: $w^T D_z x \simeq S, \quad S \sim N(\mu_S, \sigma_S^2)$
- How to approximate?

- Option 1: $\mathbf{E}_S\left[(y-\sigma(S))\right]\mathbf{E}_z\left[D_z x\right]$
- Option 2: $(y-\sigma(\mathbf{E}_S[S]))\mathbf{E}_z\left[D_z x\right]$

↓ Have closed forms but poor approximations

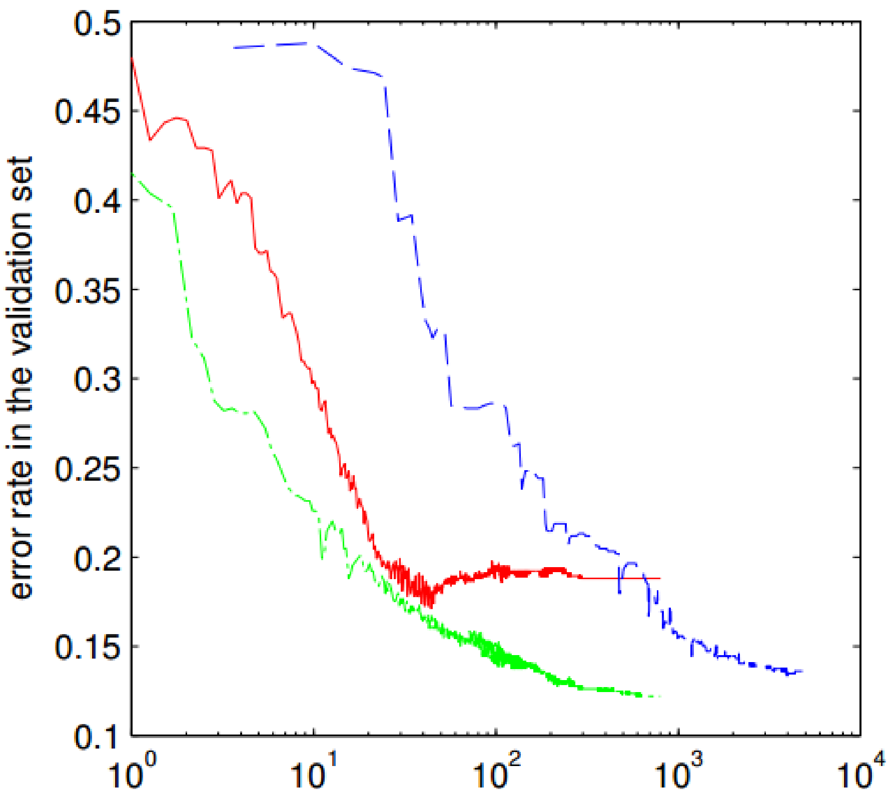
Experiment: evaluating the approximation

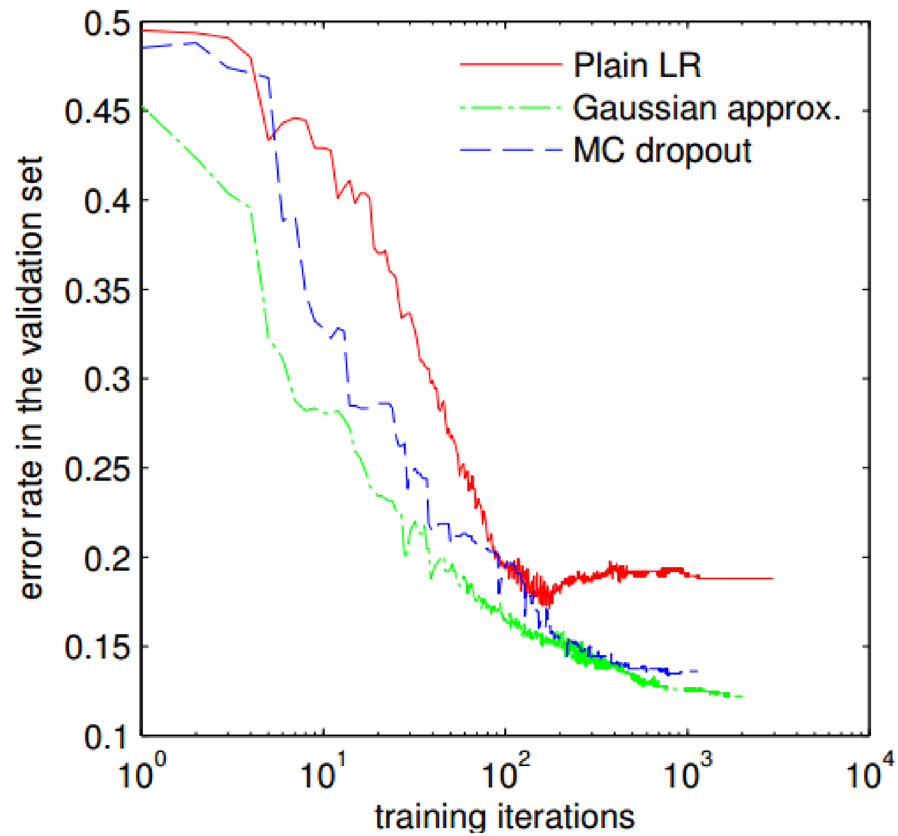
- The quality of approximation for $\Delta w_{\log}$



Experiment: Document Classification

- 20-newsgroup subtask *alt.atheism vs. religion.misc*





Experiment: Document Classification(2)

Methods\ Datasets	MR-2k	IMDB	RTs	Subj	AthR	CR	MPQA	Average
Real (MC) dropout	89.8	91.2	79.2	93.3	86.7	82.0	86.0	86.88
training time	*6400*	*6800*	*2300*	*2000*	*130*	*580*	*420*	*2700*
Gaussian dropout	89.7	91.2	79.0	93.4	87.4	82.1	86.1	86.99
training time	*240*	*1070*	*360*	*320*	*6*	*90*	*180*	*320*
Fast (closed-form) dropout	89.5	91.1	79.1	93.6	86.5	81.9	86.3	86.87
training time	*120*	*420*	*130*	*130*	*3*	*28*	*35*	*120*
plain LR	88.2	89.5	77.2	91.3	83.6	80.4	84.6	84.97
training time	*140*	*310*	*81*	*68*	*3*	*17*	*22*	*92*
Previous results								
TreeCRF(Nakagawa et al., 2010)	-	-	77.3	-	-	81.4	86.1	-
Vect. Sent.(Maas et al., 2011)	88.9	88.9	-	88.1	-	-	-	-
RNN(Socher et al., 2011)	-	-	77.7	-	-	-	86.4	-
NBSVM(Wang & Manning, 2012)	89.4	91.2	79.4	93.2	87.9	81.8	86.3	87.03
$\|\{i : x_i > 0\}\|$	788	232	22	25	346	21	4	

Fast Dropout training

- Approx: $\Delta w_{avg} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)}\left[(y-\sigma(w^T D_z x))D_z x\right]$

Fast Dropout training

- Approx: $\Delta w_{avg} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)}\left[(y - \sigma(w^T D_z x))D_z x\right]$
- By knowing:
-

Fast Dropout training

- Approx: $\Delta w_{avg} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)}\left[(y - \sigma(w^T D_z x))D_z x\right]$
- By knowing: $w^T D_z x \simeq S, \quad S \sim N(\mu_S, \sigma_S^2)$
-

Fast Dropout training

- Approx: $\Delta w_{avg} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)}\left[(y-\sigma(w^T D_z x))D_z x\right]$
- By knowing: $w^T D_z x \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$
- $\Delta w_{avg,i} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)}\left[(y-\sigma(w^T D_z x))z_i x_i\right]$

Fast Dropout training

- Approx: $\Delta w_{avg} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)} \left[(y - \sigma(w^T D_z x)) D_z x \right]$
- By knowing: $w^T D_z x \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$
- $$\Delta w_{avg,i} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)} \left[(y - \sigma(w^T D_z x)) z_i x_i \right]$$
$$= p(z_i = 1) x_i \quad \mathbf{E}_{z_{-i}|z_i=1} \left[(y - \sigma(w^T D_z x)) \right]$$

Fast Dropout training

- Approx: $\Delta w_{avg} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)}\left[(y-\sigma(w^T D_z x))D_z x\right]$
- By knowing: $w^T D_z x \simeq S, \quad S \sim N(\mu_S, \sigma_S^2)$
- $\Delta w_{avg,i} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)}\left[(y-\sigma(w^T D_z x))z_i x_i\right]$

$$= p(z_i = 1)x_i \quad \mathbf{E}_{z_{-i}|z_i=1}\left[(y-\sigma(w^T D_z x))\right]$$

Fast Dropout training

- Approx: $\Delta w_{avg} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)} \left[(y - \sigma(w^T D_z x)) D_z x \right]$
- By knowing: $w^T D_z x \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$
- $\Delta w_{avg,i} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)} \left[(y - \sigma(w^T D_z x)) z_i x_i \right]$

$$= p(z_i = 1) x_i \quad \mathbf{E}_{z_{-i} | z_i = 1} \left[(y - \sigma(w^T D_z x)) \right]$$

$$= \quad p_i \quad x_i \mathbf{E}_{z_{-i} | z_i = 1} \left[(y - \sigma(w^T D_z x)) \right]$$

Fast Dropout training

- Approx: $\Delta w_{avg} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)}\left[(y-\sigma(w^T D_z x))D_z x\right]$
- By knowing: $w^T D_z x \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$
- $$\Delta w_{avg,i} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)}\left[(y-\sigma(w^T D_z x))z_i x_i\right]$$
$$= p(z_i = 1)x_i \quad \mathbf{E}_{z_{-i}|z_i=1}\left[(y-\sigma(w^T D_z x))\right]$$
$$= \quad p_i \quad x_i \mathbf{E}_{z_{-i}|z_i=1}\left[(y-\sigma(w^T D_z x))\right]$$
$$= \quad p_i \quad x_i\left(y - \mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right]\right)$$

Fast Dropout training

- Approx: $\Delta w_{avg} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)} \left[(y - \sigma(w^T D_z x)) D_z x \right]$
- By knowing: $w^T D_z x \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$
- $$\Delta w_{avg,i} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)} \left[(y - \sigma(w^T D_z x)) z_i x_i \right]$$
$$= p(z_i = 1) x_i \quad \mathbf{E}_{z_{-i}|z_i=1} \left[(y - \sigma(w^T D_z x)) \right]$$
$$= \quad p_i \quad x_i \mathbf{E}_{z_{-i}|z_i=1} \left[(y - \sigma(w^T D_z x)) \right]$$
$$= \quad p_i \quad x_i \left(y - \mathbf{E}_{z_{-i}|z_i=1} \left[\sigma(w^T D_z x) \right] \right)$$

Fast Dropout training

- Approx: $\Delta w_{avg} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)}\left[(y-\sigma(w^T D_z x))D_z x\right]$
- By knowing: $w^T D_z x \simeq S, \quad S \sim N(\mu_S, \sigma_S^2)$
- $$\Delta w_{avg,i} = \mathbf{E}_{z;z_i \sim Bernoulli(p_i)}\left[(y-\sigma(w^T D_z x))z_i x_i\right]$$
$$= p(z_i = 1)x_i \quad \mathbf{E}_{z_{-i}|z_i=1}\left[(y-\sigma(w^T D_z x))\right]$$
$$= \quad p_i \quad x_i \mathbf{E}_{z_{-i}|z_i=1}\left[(y-\sigma(w^T D_z x))\right]$$
$$= \quad p_i \quad x_i\left(y - \mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right]\right)$$

$$\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S, \quad S \sim N(\mu_S, \sigma_S^2), \quad z_i \sim Bern(p_i)$

↓

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S,\ \ S \sim N(\mu_S, \sigma_S^2),\ \ z_i \sim Bern(p_i)$

$z_{-i} \mid z_i = 1 \Rightarrow$

$\downarrow$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S,\quad S \sim N(\mu_S, \sigma_S^2),\quad z_i \sim Bern(p_i)$

$$z_{-i} \mid z_i = 1 \Rightarrow w^T D_z x - w_i x_i z_i + w_i z_i$$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S,\ \ S \sim N(\mu_S, \sigma_S^2),\ \ z_i \sim Bern(p_i)$

$$z_{-i} \mid z_i = 1 \Rightarrow w^T D_z x - w_i x_i z_i + w_i z_i = S_i \sim N(\mu_{S_i}, \sigma_{S_i}^2)$$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S,\quad S \sim N(\mu_S, \sigma_S^2),\quad z_i \sim Bern(p_i)$

$$z_{-i} \mid z_i = 1 \Rightarrow w^T D_z x - w_i x_i z_i + w_i z_i = S_i \sim N(\mu_{S_i}, \sigma_{S_i}^2)$$

$$\mu_{S_i} =$$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S,\ \ S \sim N(\mu_S, \sigma_S^2),\ \ z_i \sim Bern(p_i)$

$$z_{-i} \mid z_i = 1 \Rightarrow w^T D_z x - w_i x_i z_i + w_i z_i = S_i \sim N(\mu_{S_i}, \sigma_{S_i}^2)$$

$$\mu_{S_i} = \mu_S + \mathbf{E}\left[-w_i x_i z_i + w_i x_i\right]$$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S,\ \ S \sim N(\mu_S, \sigma_S^2),\ \ z_i \sim Bern(p_i)$

$$z_{-i} \mid z_i = 1 \Rightarrow w^T D_z x - w_i x_i z_i + w_i z_i = S_i \sim N(\mu_{S_i}, \sigma_{S_i}^2)$$

$$\mu_{S_i} = \mu_S + \mathbf{E}\left[-w_i x_i z_i + w_i x_i\right] = \mu_S + w_i x_i (1 - p_i)$$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S,\ \ S \sim N(\mu_S, \sigma_S^2),\ \ z_i \sim Bern(p_i)$

$$z_{-i} \mid z_i = 1 \Rightarrow w^T D_z x - w_i x_i z_i + w_i z_i = S_i \sim N(\mu_{S_i}, \sigma_{S_i}^2)$$

$$\mu_{S_i} = \mu_S + \mathbf{E}\left[-w_i x_i z_i + w_i x_i\right] = \mu_S + w_i x_i (1 - p_i)$$

$$\sigma_{S_i}^2 =$$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S, \quad S \sim N(\mu_S, \sigma_S^2), \quad z_i \sim Bern(p_i)$

$$z_{-i} \mid z_i = 1 \Rightarrow w^T D_z x - w_i x_i z_i + w_i z_i = S_i \sim N(\mu_{S_i}, \sigma_{S_i}^2)$$

$$\mu_{S_i} = \mu_S + \mathbf{E}\left[-w_i x_i z_i + w_i x_i\right] = \mu_S + w_i x_i (1 - p_i)$$

$$\sigma_{S_i}^2 = \sigma_S^2 + \mathbf{Var}\left[-w_i x_i z_i + w_i x_i\right]$$

↓

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S,\ \ S \sim N(\mu_S, \sigma_S^2),\ \ z_i \sim Bern(p_i)$

$$z_{-i} \mid z_i = 1 \Rightarrow w^T D_z x - w_i x_i z_i + w_i z_i = S_i \sim N(\mu_{S_i}, \sigma_{S_i}^2)$$

$$\mu_{S_i} = \mu_S + \mathbf{E}\left[-w_i x_i z_i + w_i x_i\right] = \mu_S + w_i x_i (1 - p_i)$$

$$\sigma_{S_i}^2 = \sigma_S^2 + \mathbf{Var}\left[-w_i x_i z_i + w_i x_i\right] = \sigma_S^2 + (w_i x_i)^2 (1 - p_i) p_i$$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S,\ \ S \sim N(\mu_S, \sigma_S^2),\ \ z_i \sim Bern(p_i)$

$$z_{-i} \mid z_i = 1 \Rightarrow w^T D_z x - w_i x_i z_i + w_i z_i = S_i \sim N(\mu_{S_i}, \sigma_{S_i}^2)$$

$$\mu_{S_i} = \mu_S + \mathbf{E}\left[-w_i x_i z_i + w_i x_i\right] = \mu_S + w_i x_i (1 - p_i)$$

$$\sigma_{S_i}^2 = \sigma_S^2 + \mathbf{Var}\left[-w_i x_i z_i + w_i x_i\right] = \sigma_S^2 + (w_i x_i)^2 (1 - p_i) p_i$$

↓ $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = \mathbf{E}_{S_i}\left[\sigma(S_i)\right]$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S, \quad S \sim N(\mu_S, \sigma_S^2), \quad z_i \sim Bern(p_i)$

$$z_{-i} \mid z_i = 1 \Rightarrow w^T D_z x - w_i x_i z_i + w_i z_i = S_i \sim N(\mu_{S_i}, \sigma_{S_i}^2)$$

$$\mu_{S_i} = \mu_S + \mathbf{E}\left[-w_i x_i z_i + w_i x_i\right] = \mu_S + w_i x_i (1 - p_i)$$

$$\sigma_{S_i}^2 = \sigma_S^2 + \mathbf{Var}\left[-w_i x_i z_i + w_i x_i\right] = \sigma_S^2 + (w_i x_i)^2 (1 - p_i) p_i$$

↓ $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = \mathbf{E}_{S_i}\left[\sigma(S_i)\right]$ which could be found in closed form.

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S, \quad S \sim N(\mu_S, \sigma_S^2)$

$$z_{-i} \mid z_i = 1 \Rightarrow w^T D_z x - w_i x_i z_i + w_i z_i = S_i \sim N(\mu_{S_i}, \sigma_{S_i}^2)$$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$

$$z_{-i} \mid z_i = 1 \Rightarrow w^T D_z x - w_i x_i z_i + w_i z_i = S_i \sim N(\mu_{S_i}, \sigma_{S_i}^2)$$

$$\Delta\mu = \mu_{S_i} - \mu_S =$$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$

$$z_{-i} \mid z_i = 1 \Rightarrow w^T D_z x - w_i x_i z_i + w_i z_i = S_i \sim N(\mu_{S_i}, \sigma_{S_i}^2)$$

$$\Delta\mu = \mu_{S_i} - \mu_S =$$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$

$$z_{-i} \mid z_i = 1 \Rightarrow w^T D_z x - w_i x_i z_i + w_i z_i = S_i \sim N(\mu_{S_i}, \sigma_{S_i}^2)$$

$$\Delta\mu = \mu_{S_i} - \mu_S = w_i x_i (1 - p_i),$$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$

$$z_{-i} \mid z_i = 1 \Rightarrow w^T D_z x - w_i x_i z_i + w_i z_i = S_i \sim N(\mu_{S_i}, \sigma_{S_i}^2)$$

$$\Delta\mu = \mu_{S_i} - \mu_S = w_i x_i (1 - p_i), \quad \Delta\sigma^2 = \sigma_{S_i}^2 - \sigma_S^2 =$$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$

$$z_{-i} \mid z_i = 1 \Rightarrow w^T D_z x - w_i x_i z_i + w_i z_i = S_i \sim N(\mu_{S_i}, \sigma_{S_i}^2)$$

$$\Delta\mu = \mu_{S_i} - \mu_S = w_i x_i (1 - p_i), \quad \Delta\sigma^2 = \sigma_{S_i}^2 - \sigma_S^2 =$$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$

$$z_{-i} \mid z_i = 1 \Rightarrow w^T D_z x - w_i x_i z_i + w_i z_i = S_i \sim N(\mu_{S_i}, \sigma_{S_i}^2)$$

$$\Delta\mu = \mu_{S_i} - \mu_S = w_i x_i (1 - p_i), \quad \Delta\sigma^2 = \sigma_{S_i}^2 - \sigma_S^2 = (w_i x_i)^2 (1 - p_i) p_i$$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S, \quad S \sim N(\mu_S, \sigma_S^2)$

$$z_{-i} \mid z_i = 1 \Rightarrow w^T D_z x - w_i x_i z_i + w_i z_i = S_i \sim N(\mu_{S_i}, \sigma_{S_i}^2)$$

$$\Delta\mu = \mu_{S_i} - \mu_S = w_i x_i (1 - p_i), \quad \Delta\sigma^2 = \sigma_{S_i}^2 - \sigma_S^2 = (w_i x_i)^2 (1 - p_i) p_i$$

- S_i deviates (approximately) from S with $\Delta\mu$ and $\Delta\sigma^2$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$

$$z_{-i} \mid z_i = 1 \Rightarrow w^T D_z x - w_i x_i z_i + w_i z_i = S_i \sim N(\mu_{S_i}, \sigma_{S_i}^2)$$

$$\Delta\mu = \mu_{S_i} - \mu_S = w_i x_i (1 - p_i), \quad \Delta\sigma^2 = \sigma_{S_i}^2 - \sigma_S^2 = (w_i x_i)^2 (1 - p_i) p_i$$

- S_i deviates (approximately) from S with $\Delta\mu$ and $\Delta\sigma^2$

$$\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] =$$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S, \quad S \sim N(\mu_S, \sigma_S^2)$

$$z_{-i} \mid z_i = 1 \Rightarrow w^T D_z x - w_i x_i z_i + w_i z_i = S_i \sim N(\mu_{S_i}, \sigma_{S_i}^2)$$

$$\Delta\mu = \mu_{S_i} - \mu_S = w_i x_i (1 - p_i), \quad \Delta\sigma^2 = \sigma_{S_i}^2 - \sigma_S^2 = (w_i x_i)^2 (1 - p_i) p_i$$

- S_i deviates (approximately) from S with $\Delta\mu$ and $\Delta\sigma^2$

$$\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = \mathbf{E}_{N(\mu_S, \sigma_S^2)}\left[\sigma(S)\right] +$$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$

$$z_{-i} \mid z_i = 1 \Rightarrow w^T D_z x - w_i x_i z_i + w_i z_i = S_i \sim N(\mu_{S_i}, \sigma_{S_i}^2)$$

$$\Delta\mu = \mu_{S_i} - \mu_S = w_i x_i (1 - p_i), \quad \Delta\sigma^2 = \sigma_{S_i}^2 - \sigma_S^2 = (w_i x_i)^2 (1 - p_i) p_i$$

- S_i deviates (approximately) from S with $\Delta\mu$ and $\Delta\sigma^2$

$$\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = \mathbf{E}_{N(\mu_S,\sigma_S^2)}\left[\sigma(S)\right] + \Delta\mu \frac{\partial}{\partial\mu} \mathbf{E}_{N(\mu,\sigma_S^2)}\left[\sigma(S)\right]\Big|_{\mu=\mu_S}$$

-

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S, \quad S \sim N(\mu_S, \sigma_S^2)$

$$z_{-i} \mid z_i = 1 \Rightarrow w^T D_z x - w_i x_i z_i + w_i z_i = S_i \sim N(\mu_{S_i}, \sigma_{S_i}^2)$$

$$\Delta\mu = \mu_{S_i} - \mu_S = w_i x_i (1-p_i), \quad \Delta\sigma^2 = \sigma_{S_i}^2 - \sigma_S^2 = (w_i x_i)^2 (1-p_i) p_i$$

- S_i deviates (approximately) from S with $\Delta\mu$ and $\Delta\sigma^2$

$$\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = \mathbf{E}_{N(\mu_S,\sigma_S^2)}[\sigma(S)] + \Delta\mu \frac{\partial}{\partial\mu}\mathbf{E}_{N(\mu,\sigma_S^2)}[\sigma(S)]\Big|_{\mu=\mu_S}$$

$$+\Delta\sigma^2 \frac{\partial}{\partial\sigma^2}\mathbf{E}_{N(\mu_S,\sigma^2)}[\sigma(S)]\Big|_{\sigma^2=\sigma_S^2}$$

Fast Dropout training

- We want to: $\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = ?$
- Previously: $w^T D_z x \simeq S, \qquad S \sim N(\mu_S, \sigma_S^2)$

$$z_{-i} \mid z_i = 1 \Rightarrow w^T D_z x - w_i x_i z_i + w_i z_i = S_i \sim N(\mu_{S_i}, \sigma_{S_i}^2)$$

$$\Delta\mu = \mu_{S_i} - \mu_S = w_i x_i (1 - p_i), \quad \Delta\sigma^2 = \sigma_{S_i}^2 - \sigma_S^2 = (w_i x_i)^2 (1 - p_i) p_i$$

- S_i deviates (approximately) from S with $\Delta\mu$ and $\Delta\sigma^2$

$$\mathbf{E}_{z_{-i}|z_i=1}\left[\sigma(w^T D_z x)\right] = \mathbf{E}_{N(\mu_S, \sigma_S^2)}\left[\sigma(S)\right] + \Delta\mu \frac{\partial}{\partial\mu} \mathbf{E}_{N(\mu, \sigma_S^2)}\left[\sigma(S)\right]\Big|_{\mu = \mu_S}$$

$$+ \Delta\sigma^2 \frac{\partial}{\partial\sigma^2} \mathbf{E}_{N(\mu_S, \sigma^2)}\left[\sigma(S)\right]\Big|_{\sigma^2 = \sigma_S^2}$$

- Has closed form!