A Theoretical Framework for Deep and Locally Connected ReLU Network

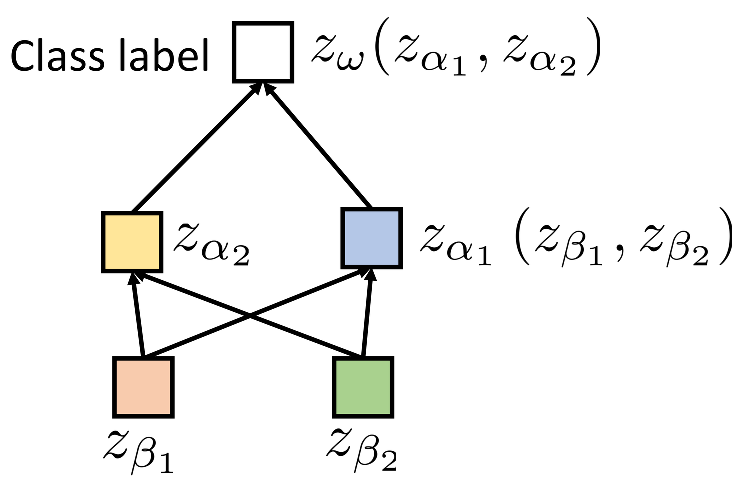
Yuandong Tian, Facebook AI Research

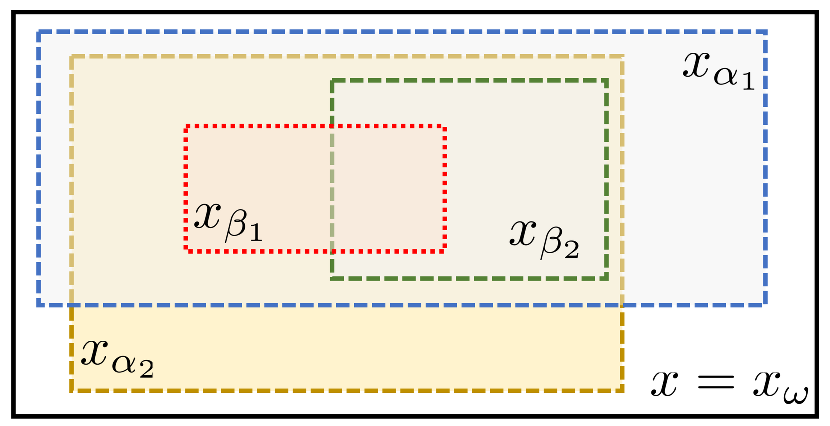
Arxiv: https://arxiv.org/abs/1809.10829

Problem Setting

Teacher (Locally connected)

Label provided by *unknown* but fixed bottom-up computational structure.
Top layer is the *class label*.

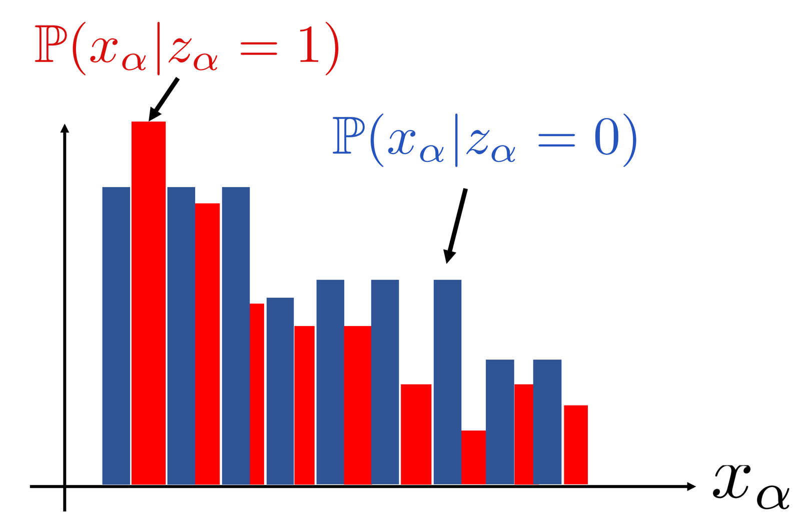




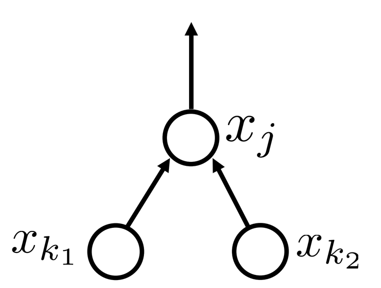
z_α Multinomial events happen in the local region of x_α

$\boxed{z_\alpha = z_\alpha(x_\alpha)}$ z is a function of x at the region

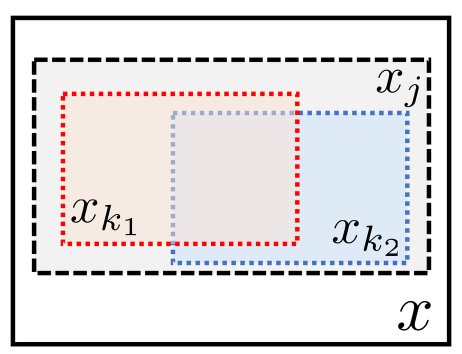
$\mathbb{P}(x_\alpha|z_\alpha)$ data distribution given a value of z_α



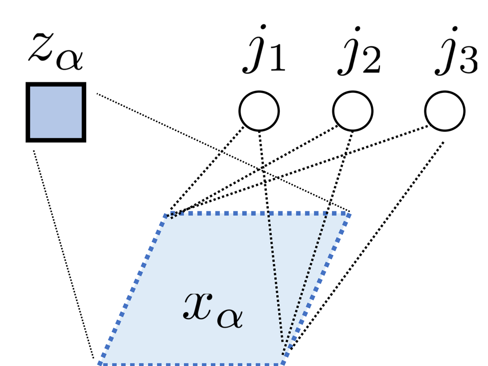
Student Network



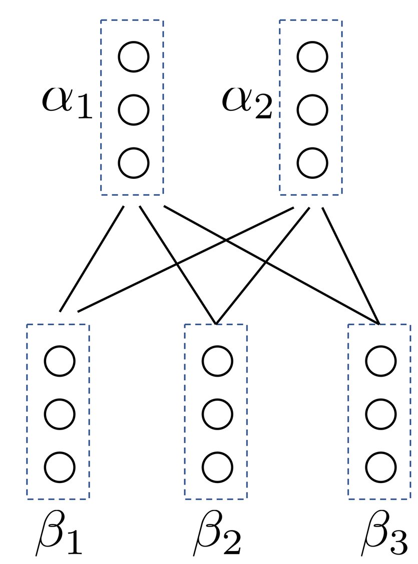
A neural network with local connectivity



image



Group nodes with the same receptive field:



Objective:
Match the output of the student network with class labels.

Goal

Student receives supervision from the highest level
Will student mimic the teacher ***at every level*** after GD?

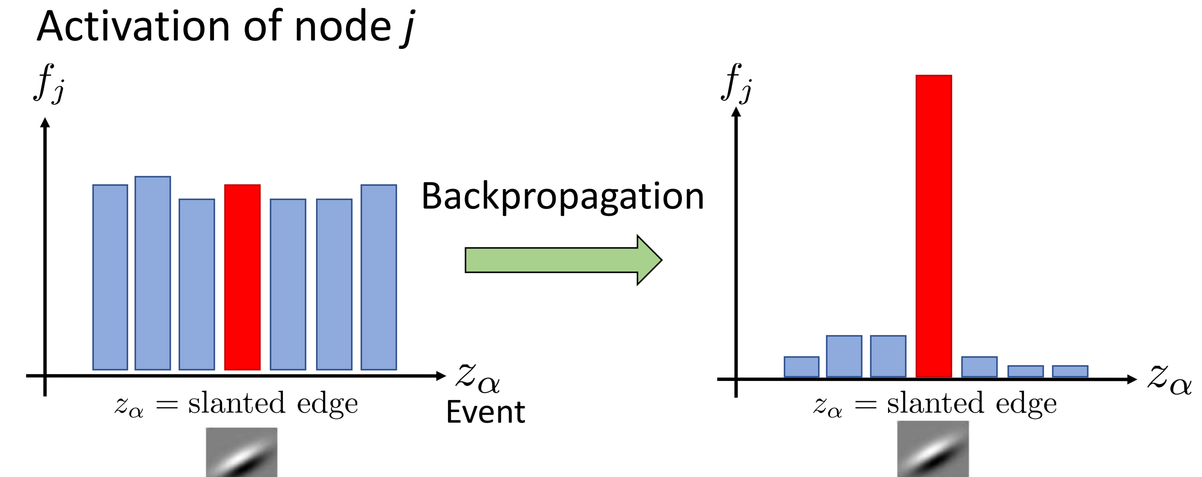
Define:

$$f_j(z_\alpha) \equiv \mathbb{E}_{x|z_\alpha}\left[f_j(x)\right]$$

What we want:

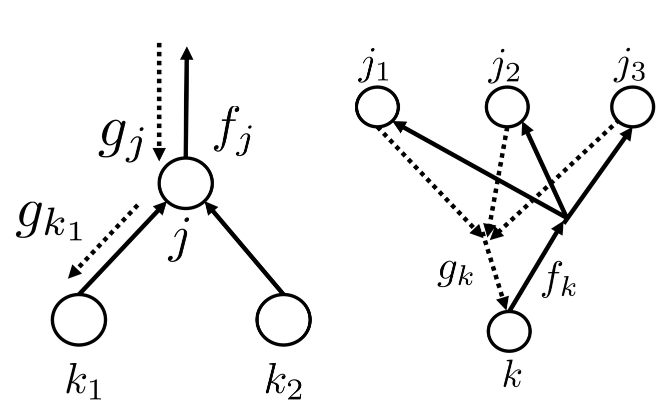
$$f_j(z_\alpha = a) > 0$$
$$f_j(z_\alpha \neq a) \approx 0$$

Activation of node *j*



Long term [Super hard]: Show propagation can achieve this goal.
First step [This paper]: Build a formulation first based on this setting and discuss about its properties

Forward and Backward Propagation in Student Network



$$f_j(x) = f'_j(x)\sum_{k\in \mathrm{ch}(j)} w_{jk} f_k(x)$$

$$g_k(x) = f'_k(x)\sum_{j\in \mathrm{pa}(k)} w_{jk} g_j(x)$$

$$\Delta w_{jk} = \mathbb{E}_x\left[f_k(x) g_j(x)\right]$$

Notation of scalar quantities (all dependent on weights)

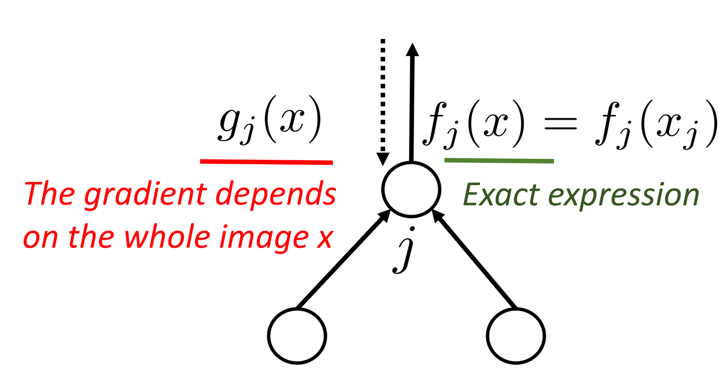
- $f_j(x)$ Activation of node j when input is x
- $f'_j(x)$ Gating of node j when input is x
- $g_j(x)$ Gradient propagated to node j, when input is x
- w_{jk} Weights connecting node j to node k

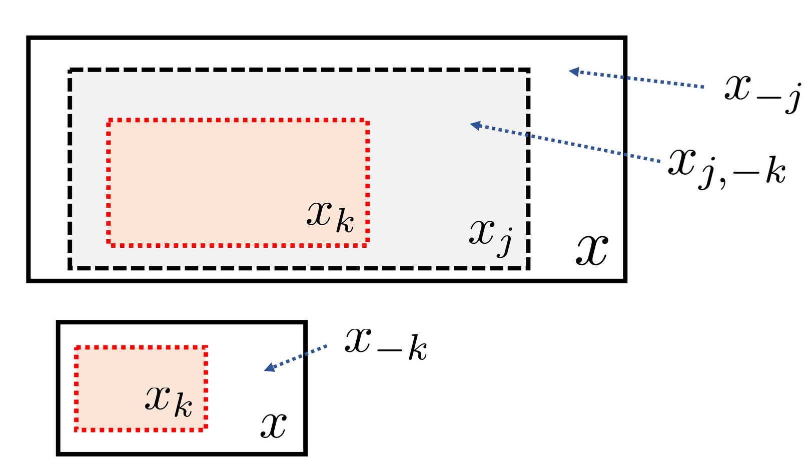
Using locality

Marginalized gradient

$$g_j(x_k) = \mathbb{E}_{x_{-k}|x_k}\left[g_j(x)\right]$$

$$g_k(x_k) = \mathbb{E}_{x_{-k}|x_k}\left[g_k(x)\right]$$





Theorem 1 (Recursive Property of marginalized gradient). $g_j(x_k) = \mathbb{E}_{x_{j,-k}|x_k}\left[g_j(x_j)\right]$

Reformulation

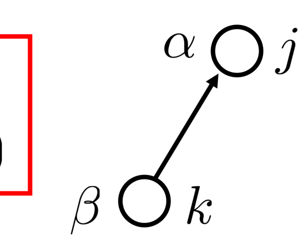
Theorem 2 [Reformulation]:

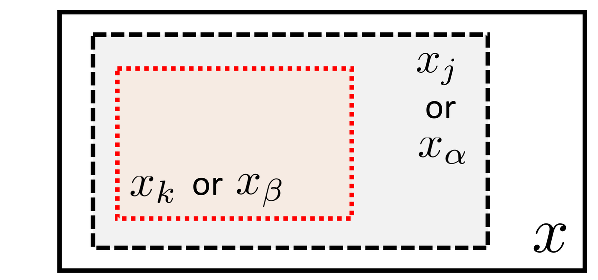
$$f_j(z_\alpha) = f'_j(z_\alpha)\sum_{k\in \mathrm{ch}(j)} w_{jk}\mathbb{E}_{z_\beta|z_\alpha}\left[f_k(z_\beta)\right]$$

$$g_k(z_\beta) = f'_k(z_\beta)\sum_{j\in \mathrm{pa}(k)} w_{jk}\mathbb{E}_{z_\alpha|z_\beta}\left[g_j(z_\alpha)\right]$$

Assumption

$$\boxed{\begin{aligned}\mathbb{P}(x_j|z_\alpha, z_\beta) &= \mathbb{P}(x_j|z_\alpha)\\ \mathbb{P}(x_k|z_\alpha, z_\beta) &= \mathbb{P}(x_k|z_\beta)\end{aligned}}$$



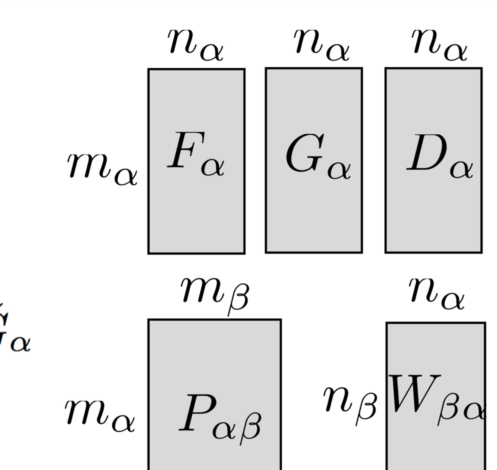


Matrix Form

$$F_\alpha = D_\alpha \circ \sum_{\beta\in \mathrm{ch}(\alpha)} P_{\alpha\beta} F_\beta W_{\beta\alpha}$$

$$\tilde{G}_\beta = D_\beta \circ \sum_{\alpha\in \mathrm{pa}(\beta)} P^T_{\alpha\beta} \tilde{G}_\alpha W^T_{\beta\alpha}$$

$$\Delta W_{\beta\alpha} = (P_{\alpha\beta}F_\beta)^T \tilde{G}_\alpha$$



n_α Number of nodes in the receptive field
m_α Number of event

$m_\alpha \to +\infty$

It becomes sample dimension

Geometrical Interpretation of Batch Norm

Forward Graph of Batch Norm

$$\hat{\mathbf{f}} = P^\perp_{\mathbf{1}}\mathbf{f}, \quad \tilde{\mathbf{f}} = \hat{\mathbf{f}}/\|\hat{\mathbf{f}}\|_{\mathrm{uni}}, \quad \bar{\mathbf{f}} = c_1\tilde{\mathbf{f}} + c_0\mathbf{1} = S(\mathbf{f})\mathbf{c}$$

$$BN(\mathbf{f}) = c_1\frac{P^\perp_{\mathbf{1}}\mathbf{f}}{\|P^\perp_{\mathbf{1}}\mathbf{f}\|_{\mathrm{uni}}} + c_0\mathbf{1} \qquad \langle f, g\rangle_{\mathrm{uni}} = \frac{1}{N}\sum_{i=1}^N f_i g_i$$

Backward Gradient from the original paper

$$\frac{\partial \ell}{\partial \hat{x}_i} = \frac{\partial \ell}{\partial y_i}\cdot\gamma$$

$$\frac{\partial \ell}{\partial \sigma^2_{\mathcal{B}}} = \sum_{i=1}^m \frac{\partial \ell}{\partial \hat{x}_i}\cdot(x_i - \mu_{\mathcal{B}})\cdot\frac{-1}{2}(\sigma^2_{\mathcal{B}}+\epsilon)^{-3/2}$$

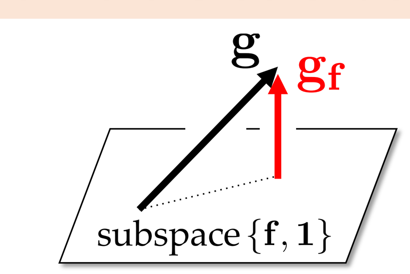
$$\frac{\partial \ell}{\partial \mu_{\mathcal{B}}} = \left(\sum_{i=1}^m \frac{\partial \ell}{\partial \hat{x}_i}\cdot\frac{-1}{\sqrt{\sigma^2_{\mathcal{B}}+\epsilon}}\right) + \frac{\partial \ell}{\partial \sigma^2_{\mathcal{B}}}\cdot\frac{\sum_{i=1}^m -2(x_i-\mu_{\mathcal{B}})}{m}$$

$$\frac{\partial \ell}{\partial x_i} = \frac{\partial \ell}{\partial \hat{x}_i}\cdot\frac{1}{\sqrt{\sigma^2_{\mathcal{B}}+\epsilon}} + \frac{\partial \ell}{\partial \sigma^2_{\mathcal{B}}}\cdot\frac{2(x_i-\mu_{\mathcal{B}})}{m} + \frac{\partial \ell}{\partial \mu_{\mathcal{B}}}\cdot\frac{1}{m}$$

$$\frac{\partial \ell}{\partial \gamma} = \sum_{i=1}^m \frac{\partial \ell}{\partial y_i}\cdot\hat{x}_i$$

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^m \frac{\partial \ell}{\partial y_i}$$

Backward Gradient of Batch Norm [Theorem 5]

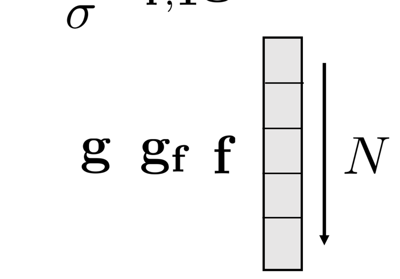


Projected Gradient

$$\mathbf{g_f} = J^{BN}(\mathbf{f})\mathbf{g} = \frac{c_1}{\sigma}P^\perp_{\mathbf{f},\mathbf{1}}\mathbf{g}$$

$$g_{c_0} = \mathbf{1}^T\mathbf{g}$$

$$g_{c_1} = \tilde{\mathbf{f}}^T\mathbf{g}$$



Batch Norm in Reformulation

Using $\mathbb{E}_x\left[f_j(x)\right] = \mathbb{E}_{z_\alpha}\left[f_j(z_\alpha)\right]$ and we can rewrite BN:

$$\mu = \mathbb{E}_{z_\alpha}\left[f_j\right], \quad \sigma^2 = \mathbb{E}_{z_\alpha}\left[(f_j(z_\alpha)-\mu)^2\right]$$

$$\mathbf{g_f} = J^{BN}(\mathbf{f})\mathbf{g} = \frac{c_1}{\sigma}\boxed{P^\perp_{\mathbf{f},\mathbf{1}}}\mathbf{g}$$

projection under new inner product $\langle f, g\rangle_{z_\alpha} = \mathbb{E}_{z_\alpha}\left[f(z_\alpha)g(z_\alpha)\right]$

Applications of the Framework

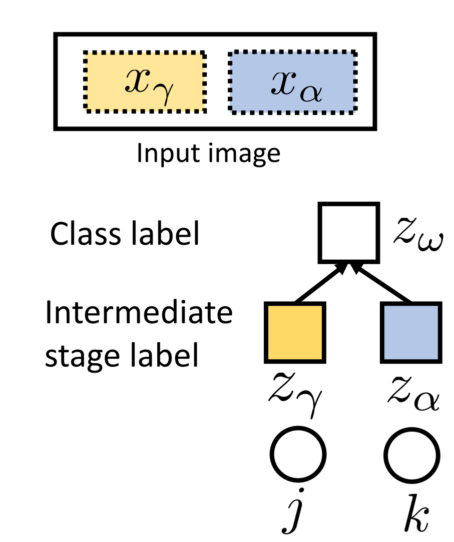
Linear versus Nonlinear

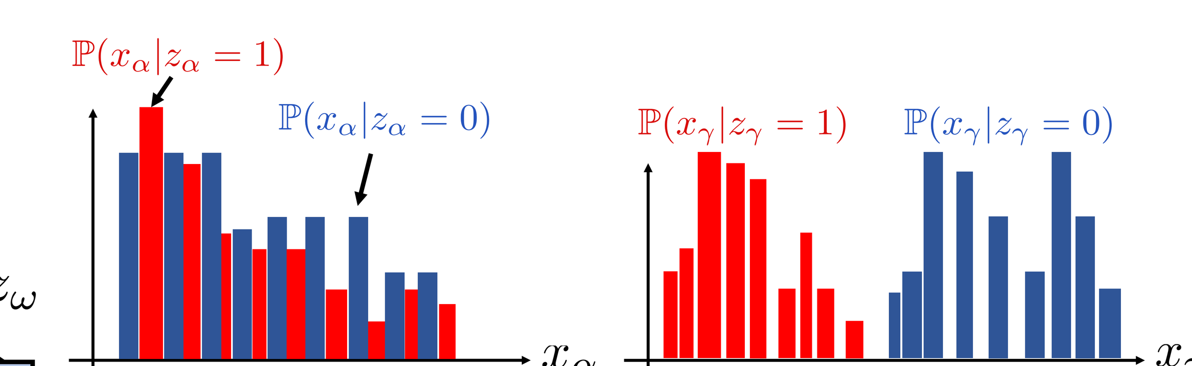
If $P_{\alpha\beta}$ is *all-vert*, $m_\alpha = n_\alpha = \mathcal{O}(\exp(\mathrm{sz}(\alpha)))$, then **[Theorem 6]**

$$\min_W Loss_{\mathrm{ReLU}}(W) = 0, \quad \min_W Loss_{\mathrm{Linear}}(W) = \mathcal{O}(\exp(\mathrm{sz}(\omega)))$$

$\mathrm{sz}(\alpha)$ Size of receptive field

Overfitting





z_α is the discriminative factor.

$\tilde{\mathbb{P}}(z_\omega = 1|z_\alpha = 1) \approx 1$
$\tilde{\mathbb{P}}(z_\omega = 1|z_\alpha = 0) \approx 0$

z_γ is a noisy factor.

$\tilde{\mathbb{P}}(z_\omega = 1|z_\gamma = 1) = 0.5+\epsilon$
$\tilde{\mathbb{P}}(z_\omega = 1|z_\gamma = 0) = 0.5-\epsilon$

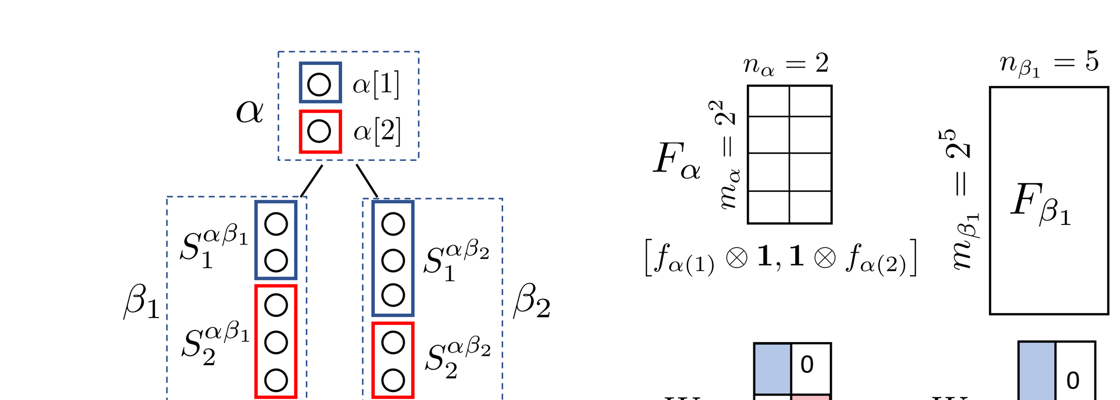
For discriminative factor

$$\Delta w_j = \mathbb{E}_{z_\alpha}\left[f_j(z_\alpha)g_0(z_\alpha)\right] \approx 0$$

For noisy factor

$$\Delta w_k = \mathbb{E}_{z_\gamma}\left[f_j(z_\gamma)g_0(z_\gamma)\right] > 0$$

Disentangled representation



$P_{\alpha\beta_1} = P_{\alpha[1]\beta_1[1,2]} \otimes P_{\alpha[2]\beta_1[3,4,5]}$
$P_{\alpha\beta_2} = P_{\alpha[1]\beta_2[1,2,3]} \otimes P_{\alpha[2]\beta_2[4,5]}$

Forward Disentanglement **[Theorem 8]**

If $P_{\alpha\beta}$ decomposes, F_α is disentangled, $W_{\beta\alpha}$ separates:
Then F_β is also disentangled.

Weight update **[Theorem 9]**

If $P_{\alpha\beta}$ decomposes, F_β is disentangled, $W_{\beta\alpha}$ separates and G_α is also disentangled, then $\Delta W_{\beta\alpha}$ also separates.

Backward disentanglement

Still in progress …