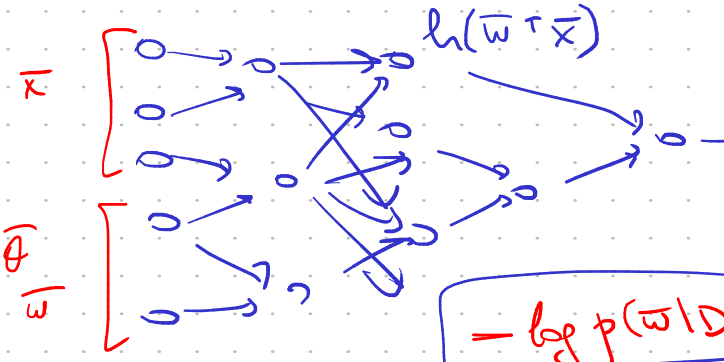
① Gradient descent



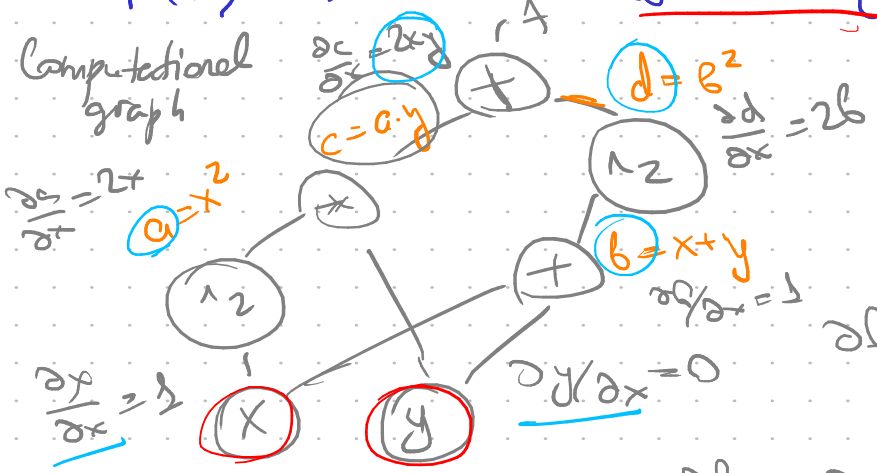
$$-\log p(D|\bar{w}) = -\sum_n \log p(\bar{x}_n|\bar{w})$$

$$-\log p(\bar{w}|D) = -\log p(D|\bar{w}) - \log p(\bar{w}) + \text{const}$$

$$\partial f/\partial x = \frac{\partial c}{\partial x} + \frac{\partial d}{\partial x} = 2xy + 2b = 2xy + 2(x+y)$$

$f(\bar{w}) \xrightarrow{\bar{w}} \min$ $\quad \bar{w} := \bar{w} - \eta \cdot \nabla_{\bar{w}} f$

Computational graph



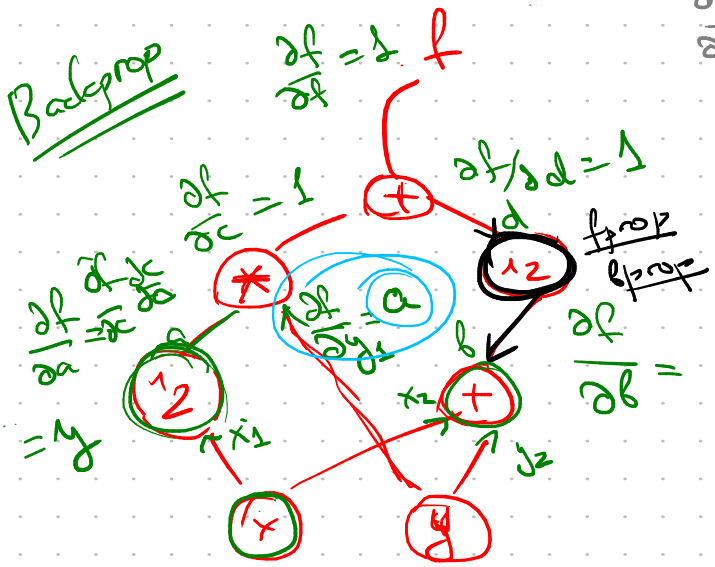
$$f(x,y) = x^2 y + (x+y)^2$$

$$\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right)$$

$\partial f/\partial x$ $\qquad (x^{(k)}, y^{(k)})$

$$\frac{\partial b}{\partial x} = \frac{\partial x}{\partial x} + \frac{\partial y}{\partial x} = 1 + 0 = 1$$

Backprop



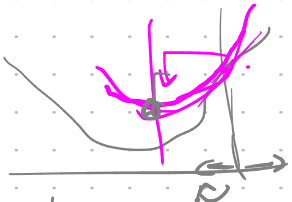
$$\frac{\partial f}{\partial b} = \frac{\partial f}{\partial d}\frac{\partial d}{\partial b} = 1 \cdot 2b = 2b$$

automatic differentiation libraries: PyTorch, TensorFlow, theano

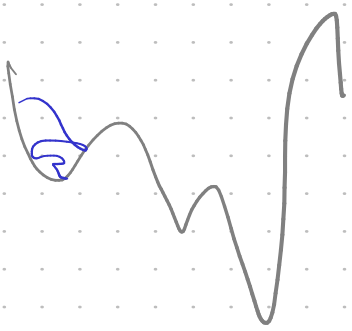
$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial a}\frac{\partial a}{\partial x} + \frac{\partial f}{\partial b}\frac{\partial b}{\partial x} = y \cdot 2x + 2b \cdot 1 = 2xy + 2b$$

$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial b}\frac{\partial b}{\partial y} + \frac{\partial f}{\partial c}\frac{\partial c}{\partial y} = 2b \cdot 1 + 1 \cdot a = a + 2b$$

GD: $\quad \bar{w}^{(k+1)} := \bar{w}^{(k)} - \eta \cdot \nabla_{\bar{w}} f \Big|_{\bar{w}^{(k)}}$



$$F(\bar{w}) \approx F(\bar{w}^{(k)}) + (\bar{w} - \bar{w}^{(k)})^T \nabla_{\bar{w}} F\Big|_{\bar{w}^{(k)}} + \underbrace{\quad}_{g_k}$$



$$+\frac{1}{2}(\bar{w}-\bar{w}^{(k)})^T H_k (\bar{w}-\bar{w}^{(k)}) \xrightarrow{\bar{w}} \min$$

$$\bar{g}_k + H_k(\bar{w}_* - \bar{w}^{(k)}) = 0 \quad \text{— second order method}$$

$$\bar{w}_* = \bar{w}^{(k)} - H_k^{-1}\bar{g}_k$$

quasi-Newton algorithms L-BFGS $H^{-1} \approx$ low-rank approx $(\bar{w}_k, \bar{g}_k)$

$$f(\bar{w}) = \sum_{n=1}^{N} f(\bar{x}_n, \bar{w}) \qquad \mathbb{E}_{\text{unif}(\bar{x}_n)}[f(\bar{x}_n, \bar{w})]$$

SGD - stochastic GD

$$F(\bar{w}) = \mathbb{E}_{q(\bar{x})}[f(\bar{x}, \bar{w})] \xrightarrow{\bar{w}} \min$$

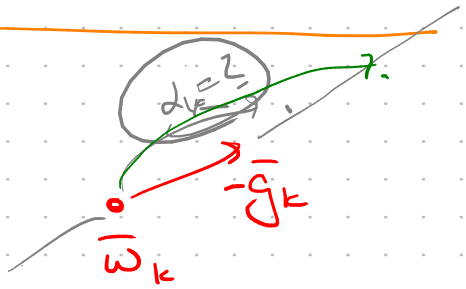
$$G(\bar{w}) = \mathbb{E}_{q(\bar{x})}[\nabla_{\bar{w}} f(\bar{x}, \bar{w})]$$

$$\mathbb{E}_{p(\bar{x})}[f(\bar{x})] \approx \frac{1}{R}\sum f(\bar{x}_i), \quad \bar{x}_i \sim p(\bar{x})$$

$$\hat{f}(\bar{w}) = \frac{1}{m}\sum_{i=1}^{m} f(\bar{x}_i, \bar{w}) \quad \text{— mini-batch}$$

$$\hat{g}(\bar{w}) = \frac{1}{m}\sum_{i=1}^{m} \nabla_{\bar{w}} f(\bar{x}_i, \bar{w})$$

SGD: $\bar{w}_{k+1} := \bar{w}_k - \eta \cdot \hat{\bar{g}}_k$



GD: $\bar{w}_{k+1} := \bar{w}_k - \alpha_k \cdot \bar{g}_k$

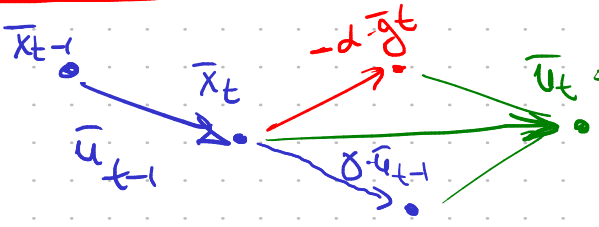
$$\varphi_k(\alpha) = F(\bar{w}_k - \alpha \cdot \bar{g}_k) \xrightarrow{\alpha} \min$$

② Вариации SGD $F(\bar{x})$ $f(x,y) = x^2 + y \cdot y^2$

$$\bar{x}_{t+1} = \bar{x}_t - \alpha \cdot \bar{g}_t$$

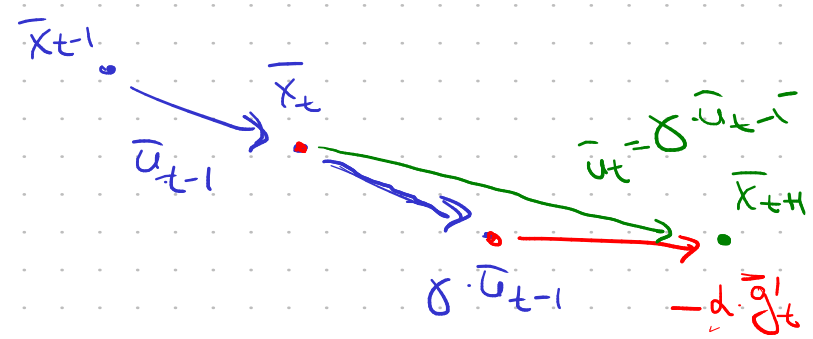
$$\bar{u}_t = \bar{x}_{t+1} - \bar{x}_t = -\alpha \cdot \bar{g}_t$$

1) SGD w/momentum $\gamma = 0.99, 0.995, 0.95$



$$\bar{u}_t = \gamma \cdot \bar{u}_{t-1} - \alpha \cdot \bar{g}_t = \nabla_x f(\bar{x}_t)$$

2) Nesterov accelerated gradients (NAG)



$$y_t = y_{t-1}\gamma + g_t(1-\gamma)$$
$$= g_t(1-\gamma) + \gamma(1-\gamma)y_{t-1} + \gamma^2(1-\gamma)y_{t-2} + \dots$$

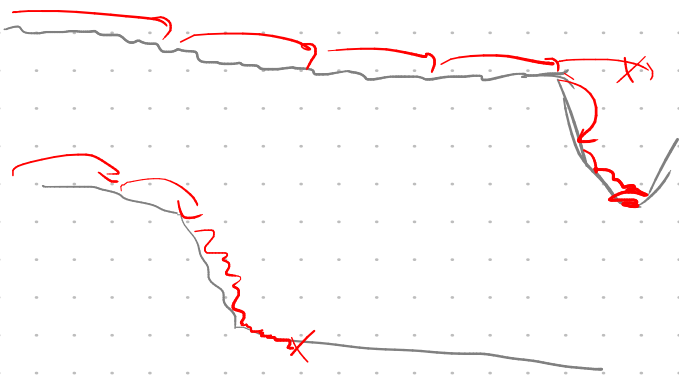
3) Adaptive SGD

- Adagrad

$G_{0,i} = 0$

$$G_{t,i} = G_{t-1,i} + g^2_{t,i}$$

$$x_{t+1,i} = x_{t,i} - \frac{\alpha}{\sqrt{G_{t,i} + \varepsilon}} \cdot g_{t,i}$$



- RMSprop

$$G_{t,i} := G_{t-1,i} \cdot \gamma + (1-\gamma) \cdot g^2_{t,i}$$

- Adadelta

$$\bar{x} := \bar{x} - \alpha \cdot \bar{g}$$

$$\bar{x} := \bar{x} - H^{-1}\bar{g}$$

$$x_{t+1,i} = x_{t,i} - \alpha \cdot \frac{\sqrt{R_{t,i} + \varepsilon}}{\sqrt{G_{t,i} + \varepsilon}} g_{t,i}, \qquad R_{t,i} = \beta R_{t,i-1} + (1-\beta) \cdot u^2_{t,i}$$

- Adam (2014)

$$G_{t,i} = \beta_2 G_{t-1,i} + (1-\beta_2) g^2_{t,i}$$

$$x_{t+1,i} = x_{t,i} - \frac{\alpha}{\sqrt{G_{t,i} + \varepsilon}} \cdot m_{t,i}$$

$$m_{t,i} = \beta_1 m_{t-1,i} + (1-\beta_1) g_{t,i}$$

$$\beta_1 = 0.9, \quad \beta_2 = 0.999, \quad \varepsilon = 10^{-7}$$

- AdamW

weight decay

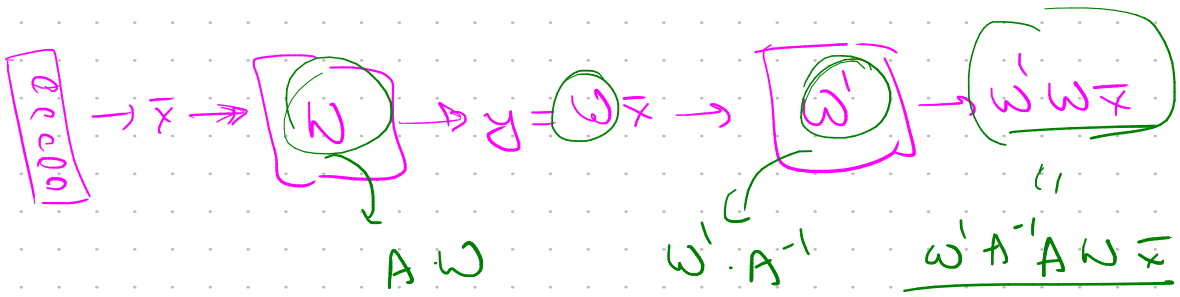
$$\bar{\omega} := (1-\eta)\bar{\omega} - \alpha \cdot \nabla_{\bar{\omega}} L$$

SGD

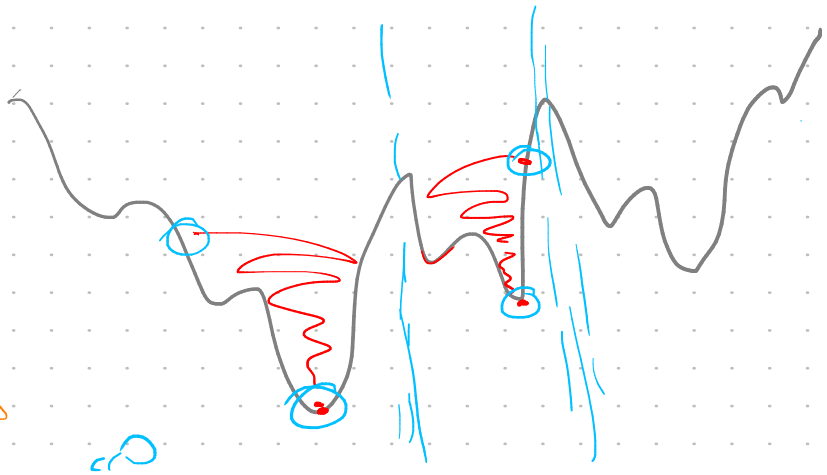
$$\bar{\omega} - \eta \cdot \bar{\omega} - \alpha \cdot \nabla_{\bar{\omega}} L(\bar{\omega}) =$$

$$= \bar{\omega} - \alpha \cdot \nabla_{\bar{\omega}} \left[L(\bar{\omega}) + \frac{\eta}{2\alpha} \cdot \|\bar{\omega}\|^2 \right]$$

(4) Dropout



(5) Weight Initialization



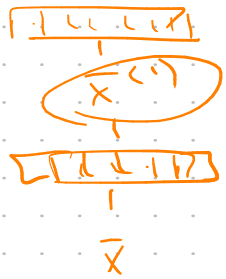
pretraining

unsupervised pretraining

$$y = \bar{w}^T\bar{x} = \sum_i w_i x_i \qquad y_i$$

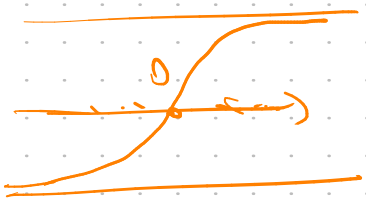
$$Var[y] = \sum_i Var[y_i]$$

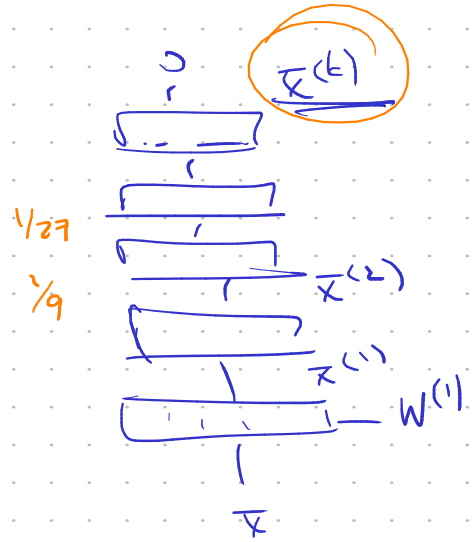
$$Var[y_i] = E[x_i]^2 Var[w_i] + E[w_i]^2 Var[x_i] + Var[w_i]Var[x_i]$$



$$E[y] = 0$$

$$\bar{x}^{(1)} = h(\bar{y})$$





$$Var[y_i] = Var[w_i] \cdot Var[x_i]$$

$$Var[y] = n \cdot Var[w_i] \cdot Var[x_i]$$

$$n \cdot Var[w_i] \approx 1 \qquad Var[w_i] \approx 1/n$$

$$Var[w_i] = O\left(\frac{1}{n}\right)$$

$$w_i \sim Unif\left(\left[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right]\right)$$

Xavier Glorot (Bengio)

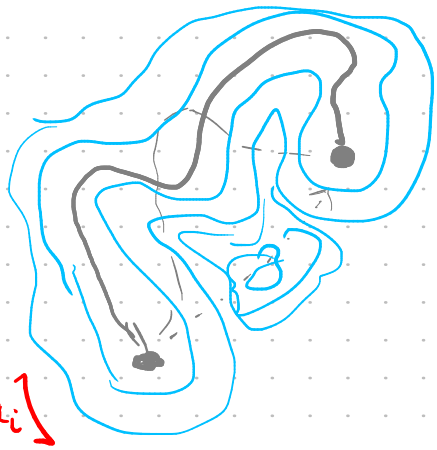
$$Var\left(unif([a,b])\right) = \frac{(b-a)^2}{12}$$

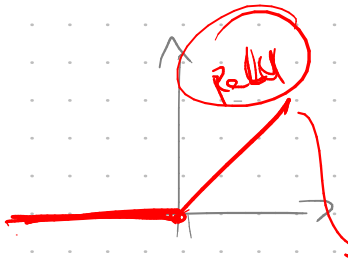
$$Var\left[unif\left(\left[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right]\right)\right] = \frac{(2/\sqrt{n})^2}{12} = \frac{1}{3n}, \quad n \cdot Var[w_i] \approx \frac{1}{3}$$

Xavier init

$w_i \sim Unif\left(\left[-\sqrt{\frac{3}{n}}, \sqrt{\frac{3}{n}}\right]\right)$

$N\left(0, \frac{1}{n}\right)$

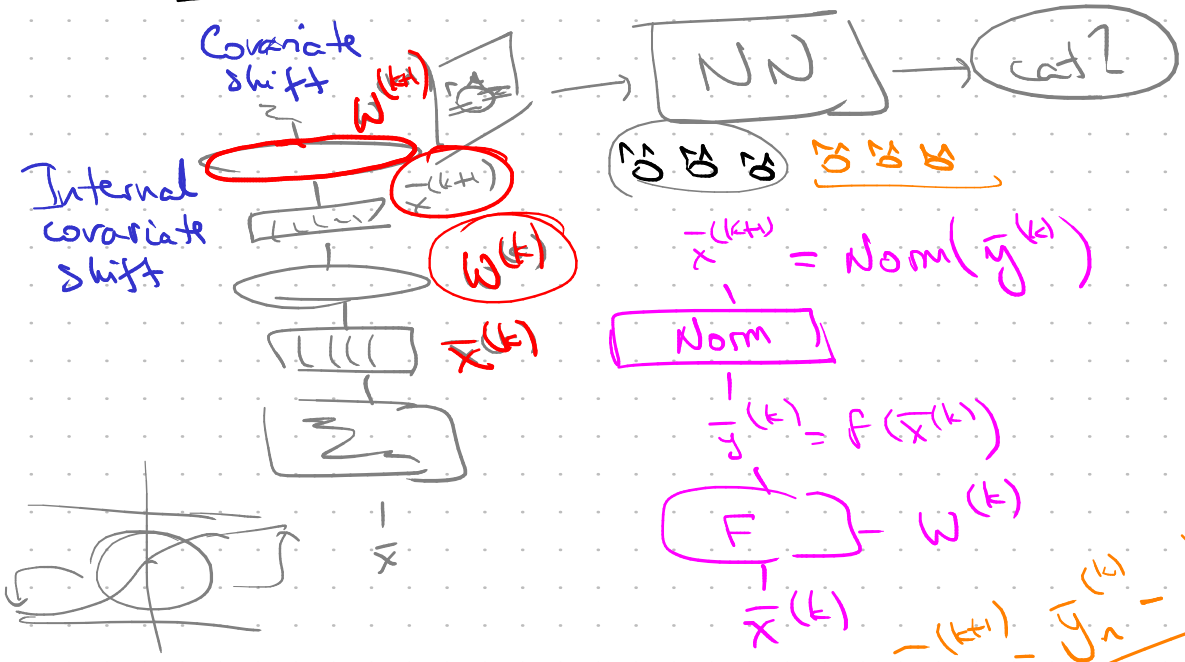




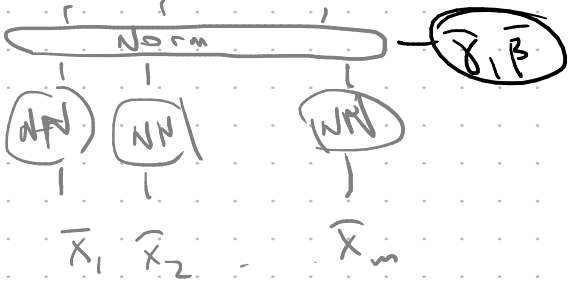
$Var[y_i] = E[x_i]^2 Var[w_i] + Var[w_i] Var[x_i]$

$\rightarrow Var[w_i] = \frac{2}{n}$

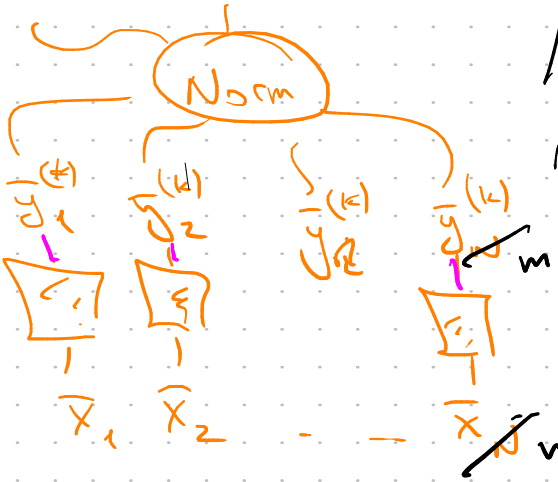
(6) Batch normalization

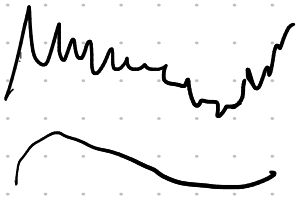


$$\bar{x}^{(k+1)} = \frac{\bar{y}^{(k)} - E_b[\bar{y}^{(k)}]}{\sqrt{Var_b[\bar{y}^{(k)}]}}$$



$$x_n^{(k+1)} = \frac{y_n^{(k)} - \frac{1}{m}\sum_{i=1}^{m} y_i^{(k)}}{\sqrt{Var[\ldots]}}$$





$$BN(\bar{y}^{(k)}) = \frac{\bar{y}^{(k)} - E_m[\bar{y}^{(k)}]}{\sqrt{Var_m[\bar{y}^{(k)}]}} \odot \bar{\gamma} + \bar{\beta}$$