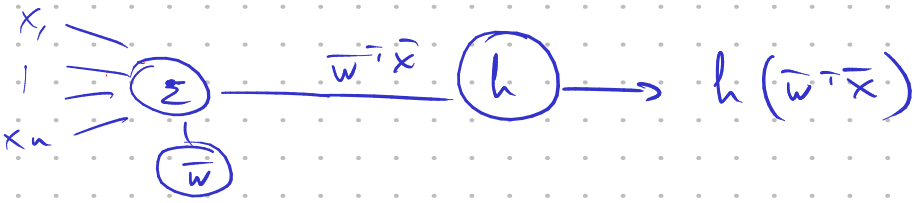
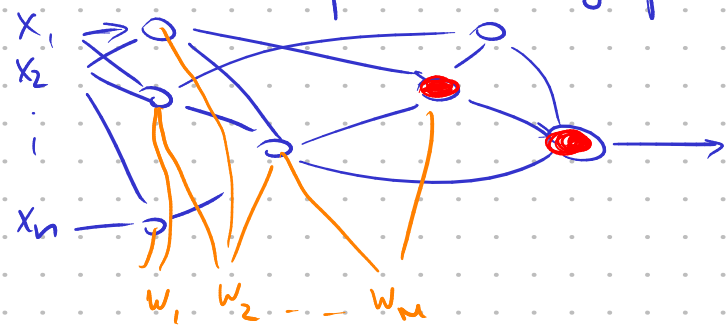


Computational graph



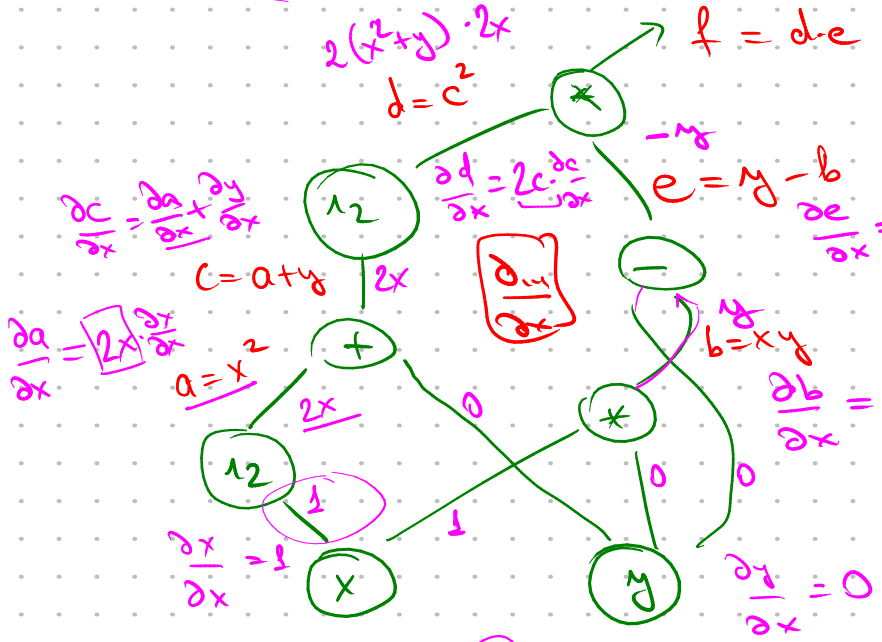
$L \to \min$

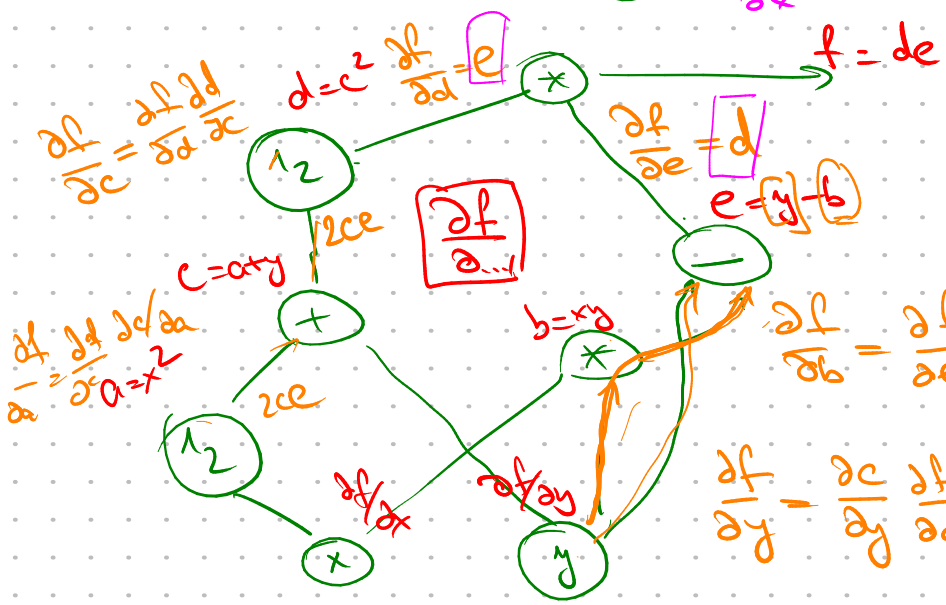
$\log p(D|\theta) \to \max$

$\bar{w} := \bar{w} - \eta \nabla_{\bar{w}} L$

$f = (x^2+y)^2 \cdot (y - xy)$

$\partial f/\partial x$ $\partial f/\partial y$





fprop
bprop

- PyTorch
- Torch
- TensorFlow
- MxNet
- theano
- ...

$$w := w - \eta \nabla_w L$$

$$L = \frac{1}{N} \sum_{n=1}^{N} l(y_n, F(x_n))$$

$\nabla_w l$

$$H = \nabla_w \nabla_w L$$

quasi-Newton methods L-BFGS

$q(y) = \text{Unif}(D)$

Stochastic gradient descent

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

$$G(x) = \nabla_x F(x)$$

mini-batch

$$\hat{f}(x) = \frac{1}{m} \sum_{i=1}^{m} f(x, y_i), \quad \hat{g}(x) = \frac{1}{m} \sum_{i=1}^{m} \nabla_x f(x, y_i)$$

$$x_{k+1} = x_k - \alpha_k \hat{g}_k \qquad g_k = \mathbb{E}\hat{g}_k = \nabla F(x_k)$$

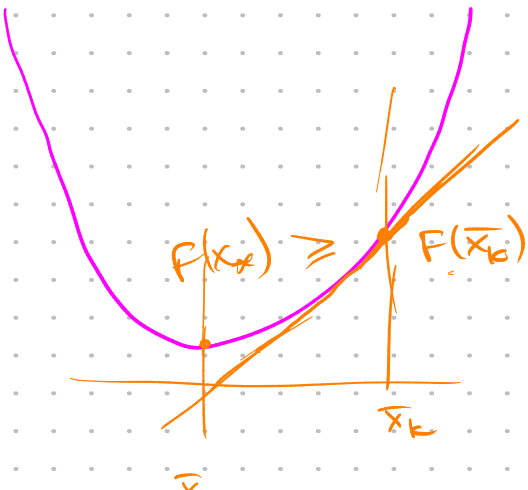
$$\|x_{k+1} - x_*\|^2 = \|x_k - \alpha_k \hat{g}_k - x_*\|^2 =$$
$$= \|x_k - x_*\|^2 - 2\alpha_k \hat{g}_k^T (x_k - x_*) + \alpha_k^2 \|\hat{g}_k\|^2$$

$$\mathbb{E}\|x_{k+1} - x_*\|^2 = \mathbb{E}\|x_k - x_*\|^2 - 2\alpha_k g_k^T (x_k - x_*) + \alpha_k^2 \mathbb{E}\|\hat{g}_k\|^2$$

$$F(x_*) \geq F(x_k) + g_k^T (x_* - x_k)$$

$$\alpha_k F(x_*) \geq \alpha_k \left(F(x_k) + g_k^T (x_* - x_k) \right)$$

$$\alpha_k g_k^T (x_k - x_*) \geq \alpha_k (F(x_k) - F(x_*))$$



$$\mathbb{E}\|x_{k+1} - x_*\|^2 - \mathbb{E}\|x_k - x_*\|^2 \leq$$
$$\leq -2\alpha_k (F(x_k) - F(x_*)) + \alpha_k^2 \mathbb{E}\|\hat{g}_k\|^2$$

$$\alpha_k (F(x_k) - F(x_*)) \leq \frac{1}{2}\mathbb{E}\|x_k - x_*\|^2 - \frac{1}{2}\mathbb{E}\|x_{k+1} - x_*\|^2 + \frac{1}{2}\alpha_k^2 \mathbb{E}\|\hat{g}_k\|^2$$

$$\sum_{i=0}^{k} \alpha_i \left(\mathbb{E} F(\bar{x}_i) - F(\bar{x}_*) \right) \le \frac{1}{2} \mathbb{E} \|\bar{x}_0 - \bar{x}_*\|^2 - \frac{1}{2} \mathbb{E} \|\bar{x}_{k+1} - \bar{x}_*\|^2 + \frac{1}{2} \sum_i \alpha_i^2 \mathbb{E} \|\hat{g}_i\|^2$$

$$\mathbb{E} F\left(\frac{\sum \alpha_i \bar{x}_i}{\sum \alpha_i} \right) \le \frac{\sum \alpha_i \mathbb{E} F(\bar{x}_i)}{\sum \alpha_i} - \sum \alpha_i F(x_*)$$

$$\mathbb{E} F\left(\frac{\sum \alpha_i \bar{x}_i}{\sum \alpha_i} \right) - F(x_*) \le \frac{\sum \alpha_i \left(\mathbb{E} F(\bar{x}_i) - F(\bar{x}_*) \right)}{\sum \alpha_i} \le \frac{}{\sum \alpha_i}$$

$$\|\bar{x}_0 - \bar{x}_*\| \le R, \quad \mathbb{E} \|\hat{g}_k\|^2 \le G^2$$

$$\mathbb{E} F(\hat{\bar{x}}_k) - F(\bar{x}_*) \le \frac{R^2 + G^2 \sum_i^k \alpha_i^2}{2 \sum_{i=0}^k \alpha_i} \le \text{Konst}$$

SGD

$\to \infty$

$\alpha_i = h$

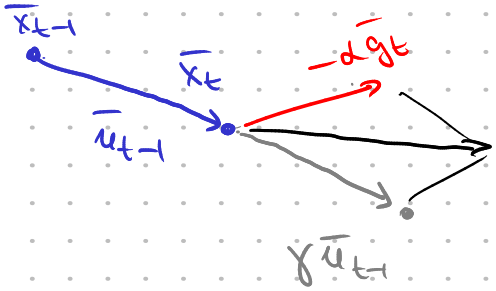
$$\dots \le \frac{R^2}{2hk} + \frac{G^2 h}{2}$$

$\to 0$, $k \to \infty$

SGD

$\bar{x}_t \xrightarrow{-\nabla F = \bar{g}_t} \bar{x}_{t+1} = \bar{x}_t - \alpha \bar{g}_t$

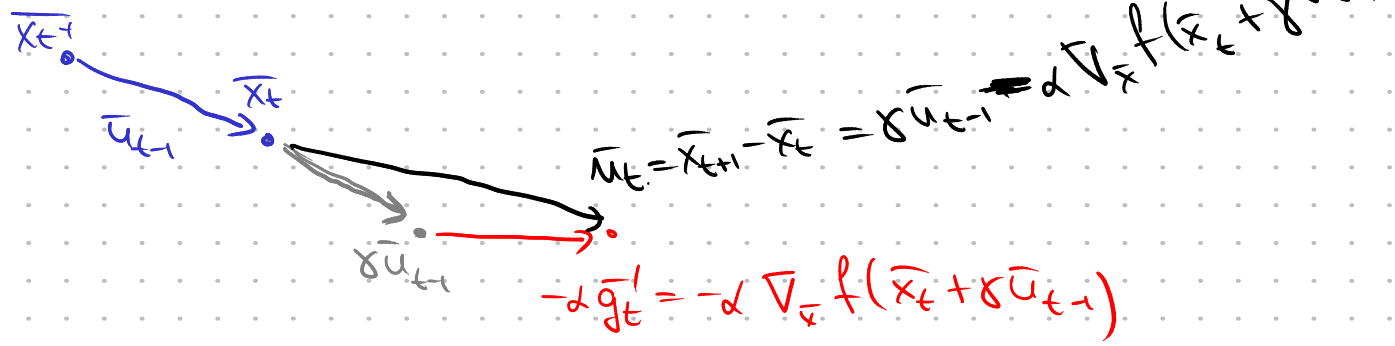
SGD with momentum



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

$\gamma = 0.9, 0.99, 0.995$

Nesterov accelerated gradient (NAG)



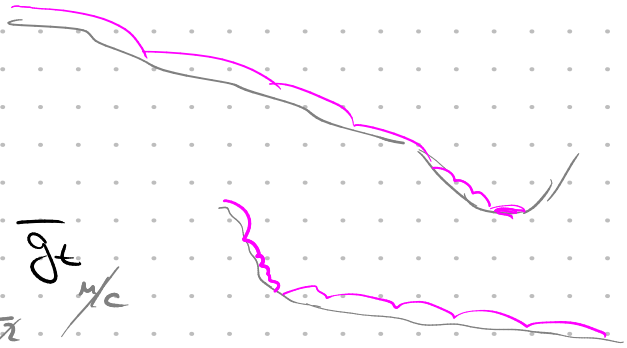
Adaptive SGD

Adagrad

$$G_0 = 0$$

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

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



RMSprop

$$G_{t,i} = \gamma G_{t-1,i} + (1-\gamma) g_{t,i}^2$$

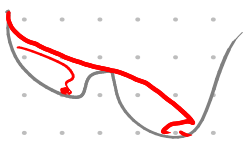
Adadelta

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

$$\bar{x}_{t+1} = \ldots H^{-1}\bar{g}_t$$

$$\bar{x}_{t+1} = \bar{x}_t - \alpha \cdot \frac{\sqrt{R_t + \varepsilon}}{\sqrt{G_t + \varepsilon}}\bar{g}_t$$

$$R_{t,i} = \rho R_{t-1,i} + (1-\rho) u_{t,i}^2$$



Adam

$$\bar{x}_{t+1} = \bar{x}_t - \frac{\alpha}{\sqrt{G_t + \varepsilon}} \cdot \bar{m}_t$$

где

$$\beta_1 = 0.9$$

$$\beta_2 = 0.999$$

$$\varepsilon = 10^{-8}$$

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

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

Nadam

AdamW