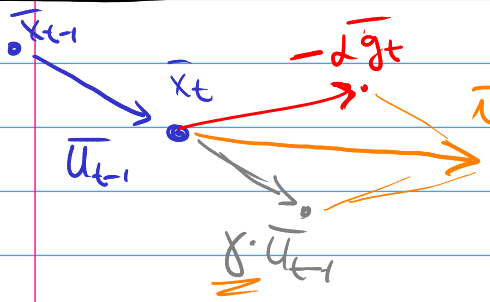
SGD

$\hat{g} = \frac{1}{m}\Sigma \dots$

$\bar{x}_t \quad -\bar{g}_t = -\nabla f(\bar{x}_t)$

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

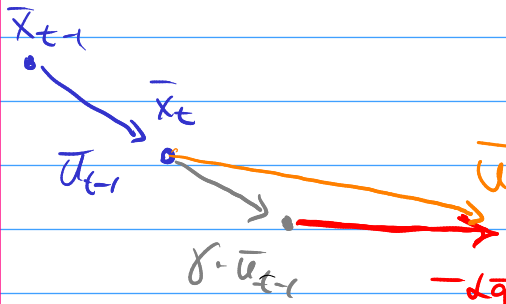
SGD with Momentum

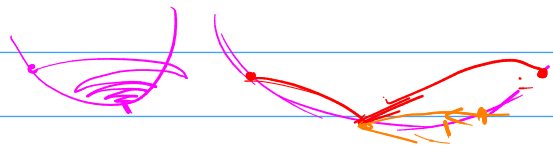


$\bar{u}_t = \gamma \bar{u}_{t-1} - \alpha \bar{g}_t$

$\gamma = 0.9,\ 0.95,\ 0.99$

Nesterov accelerated gradients (NAG)





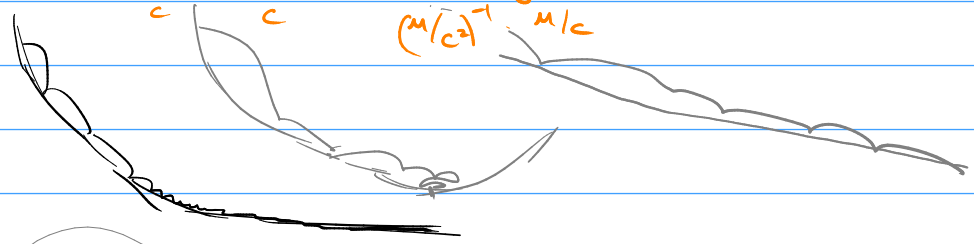
$\bar{u}_{t+1} = \gamma \bar{u}_{t-1} - \alpha \nabla_{\bar{x}} f(\bar{x}_t + \gamma \bar{u}_{t-1})$

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

Adaptive SGD

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

$c \quad c \quad (M/c^2)^{-1} \quad M/c$



Adagrad

α

$\bar{G}_0 = 0$

$G_{t,i} = G_{t-1,i} + g_{t,i}^2 \quad , \quad \bar{g}_t = \nabla_{\bar{x}} f(\bar{x}_t)$

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

$c = c - \frac{1}{\sqrt{(M/c^2)^2}} \cdot M/c$

$c = c - 1 \cdot M/c$

RMSprop

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

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

Adadelta

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

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

Adam

$$\bar{x}_{t+1} = \bar{x}_t - \frac{\alpha}{\sqrt{G_t + \varepsilon}} \bar{m}_t, \quad \text{где}$$

$$G_t = \beta_2 G_{t-1} + (1-\beta_2)\,\bar{g}^2_t$$

$$\bar{m}_t = \beta_1 \bar{m}_{t-1} + (1-\beta_1)\,\bar{g}_t$$

$\beta_1 = 0.9$

$\beta_2 = 0.999$

$\varepsilon = 10^{-8}$

AMSGrad

$\sqrt{G'_t + \varepsilon}$, где $G'_t = \max(G_t, G'_{t-1})$

Nadam

AdamW

1980-е: weight decay

$$\bar{w}_{k+1} = (1-\beta)\,\bar{w}_k + \alpha \nabla_{\bar{w}} F(\bar{w}_k) =$$

$$= \bar{w}_k - \alpha\left(\nabla_w f + \frac{\beta}{\alpha}\bar{w}_k\right)$$

$$= \bar{w}_k - \alpha \cdot \nabla_w \left(f + \frac{\beta}{2\alpha}\bar{w}^T\bar{w}\right)$$

L_2-регуляр.