Gradient descent

$f(\bar{x})$

$\bar{x}_0$,

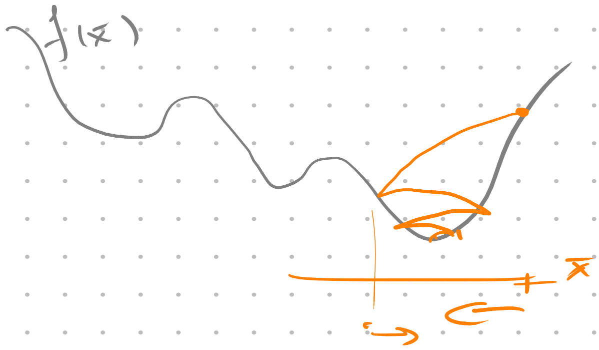
weight init

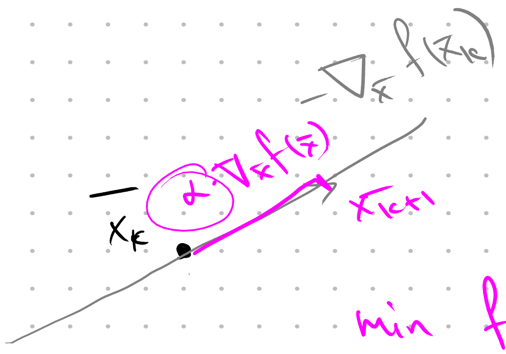
$$\bar{x}_k = \bar{x}_{k-1} - \alpha \nabla_{\bar{x}} f(\bar{x}_k)$$

$\alpha = \alpha_k \to 0$

$$\alpha_k = \alpha_0 \left(1 - \frac{k}{T}\right)$$

$$\alpha_k = \alpha_0 e^{-k/T}$$



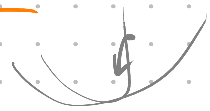


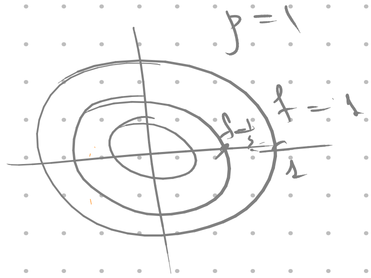
$$\min_{\alpha} \underbrace{f(\bar{x}_k - \alpha \nabla_{\bar{x}} f(\bar{x}))}_{= \varphi_k(\alpha)}$$

– 1-dim. optim.

Armijo rule $\quad \varphi_k(\alpha) \le \varphi_k(0) + c \cdot \alpha \cdot \varphi_k'(0)$, $c \in (0, \frac{1}{2})$

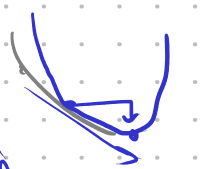
$$f(x, y) = x^2 + p \cdot y^2$$

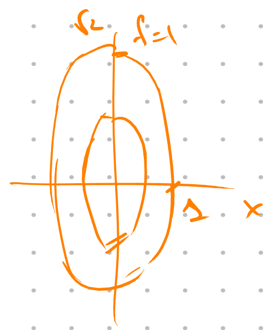




Second order methods

Newton's method





$p = \frac{1}{2}$

$$H_k = \nabla^2_{\bar{x}} f(\bar{x}_k) = \left(\cdots \frac{\partial^2 f}{\partial x_i \partial x_j} \cdots \right)$$

hessian

$$\bar{x}_{k+1} = \bar{x}_k - \alpha_k H_k^{-1} \cdot \nabla_{\bar{x}} f(\bar{x}_k)$$

quasi-Newton methods

L-BFGS $\quad H^{-1} \approx \ldots \nabla_x f(\bar{x}_k) \ldots$

Stochastic gradient descent

$$\underline{F(\bar{x})} = \underbrace{\frac{1}{N}\sum_{n=1}^{N} f(\bar{x}, \underline{\bar{y}_n})} \xrightarrow[\bar{x}]{} \min$$

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

$= \text{uniform}(D)$

$M \ll N$

$$F(\bar{x}) \approx \frac{1}{M}\sum_{m=1}^{M} f(\bar{x}, \bar{y}_m)$$

mini-batches

$$\bar{g}_k(\bar{x}) \approx \hat{g}_k = \frac{1}{M}\sum_{m=1}^{M} \nabla_{\bar{x}} f(\bar{x}, \bar{y}_m)$$

SGD with α_k , $\bar{x}_k$, $\bar{x}_{opt}$

$\|\bar{x}_0 - \bar{x}_{opt}\| \le R$

$\mathbb{E}\,\|\hat{g}_k\|^2 \le G^2$

$$\underline{\mathbb{E}\,F(\bar{x}_k) - F(\bar{x}_{opt})} \le \frac{R^2 + G^2 \cdot \sum_{i=0}^{k} \alpha_i^2}{2\sum_{i=0}^{k} \alpha_i} \xrightarrow[t\to\infty]{?} 0$$

① $\alpha_i = h = \text{const}$

$$\le \frac{R^2 + G^2 \cdot k \cdot h^2}{2 \cdot k \cdot h} = \frac{R^2}{2kh} + \frac{1}{2}G^2 h$$

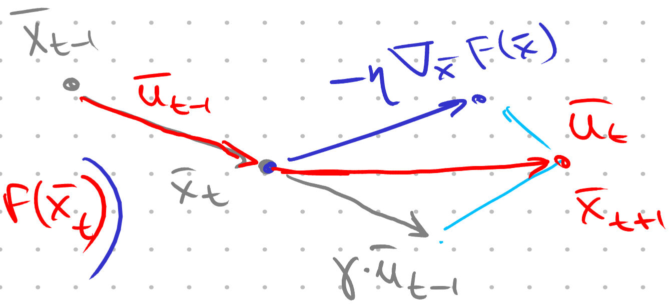
$\frac{R^2}{2kh} \xrightarrow[t\to\infty]{} 0$, $\frac{1}{2}G^2 h$ – const

② SGD converges if $\sum_0^{\infty} \alpha_i = \infty$ and $\sum_0^{\infty} \alpha_i^2 < \infty$

$$\alpha_i = \frac{1}{i}$$

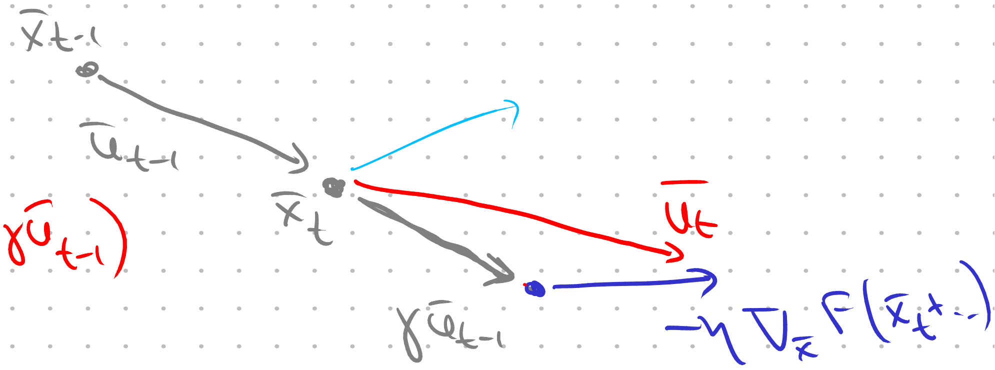
---

① Momentum



$$\bar{x}_{t+1} = \bar{x}_t + \underbrace{\left(\gamma \cdot \bar{u}_{t-1} - \eta\nabla_{\bar{x}} F(\bar{x}_t)\right)}_{= \bar{u}_t}$$

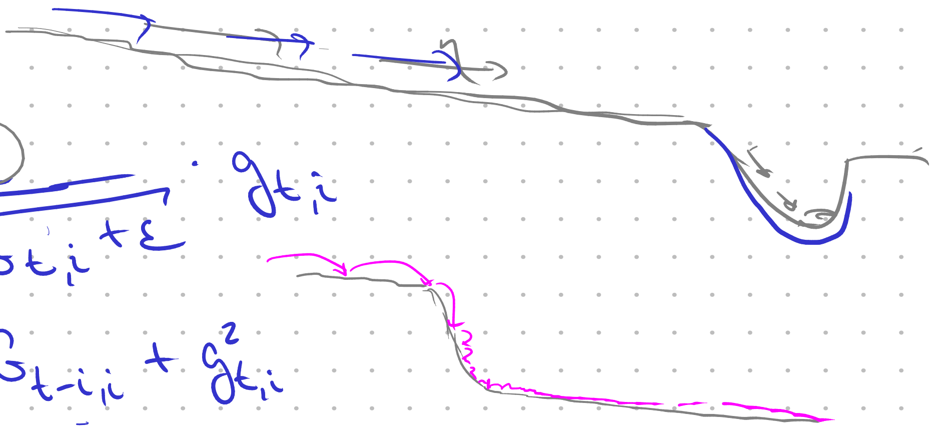
2 Nesterov Accelerated Gradients



$$\bar{u}_t = \gamma \bar{u}_{t-1} - \eta \nabla_x F(\bar{x}_t + \gamma \bar{u}_{t-1})$$

$$\bar{x}_{t+1} = \bar{x}_t + \bar{u}_t$$

3 AdaGrad - adaptive gradient

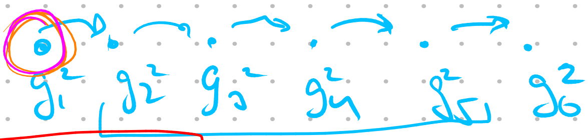


$$\bar{g}_t = \nabla_{\bar{x}} F(\bar{x}_t)$$

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

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

4 RMSprop



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

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

$$= (1-\rho) g_{t,i}^2 + (1-\rho)\rho\, g_{t-1,i}^2 + \dots + (1-\rho)\rho^k g_{t-k,i}^2 + \dots$$

5 Adadelta

$F(x)$ - meters $\qquad \nabla_x F \quad m/s$

x - seconds

$$\text{sec} = \text{sec} + \frac{\overset{\text{sec}}{\text{const}}}{\sqrt{m^2/s^2}} \cdot m/s = \text{sec} + \text{const}$$

sec = sec + const · m/s

$$\bar{x}_{k+1} = \bar{x}_k + \alpha \cdot H_k^{-1} \bar{g}_k$$

sec   sec $(m/s^2)^{-1}$ m/s

$$u_t = \rho\, u_{t-1} + (1-\rho) \cdot \Delta x_{t,i}^2$$

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

---

⑥ Adam (Kingma, Ba, 2014)

$$m_t = \underline{\beta_1} m_{t-1} + (1-\beta_1) g_t$$

$$u_t = \frac{\eta}{\sqrt{v_t + \varepsilon}} \cdot m_t \,, \quad v_t = \underline{\beta_2} v_{t-1} + (1-\beta_2) g_t^2$$

$$\beta_1 = 0.9 \,, \quad \beta_2 = 0.999 \,, \quad \varepsilon = 10^{-8}$$