Computational Intelligence

Samaneh Hosseini

Isfahan University of Technology

Outline

- Optimization Algorithms
  - Mini-batch Gradient Descent
  - Understanding Mini-batch Gradient Descent
  - Exponentially Weighted Averages
  - Understanding Exponentially Weighted Averages
  - Bias Correction in Exponentially Weighted Averages
  - Gradient Descent with Momentum
  - RMSprop

Optimization Algorithms: Understanding Exponentially Weighted Averages

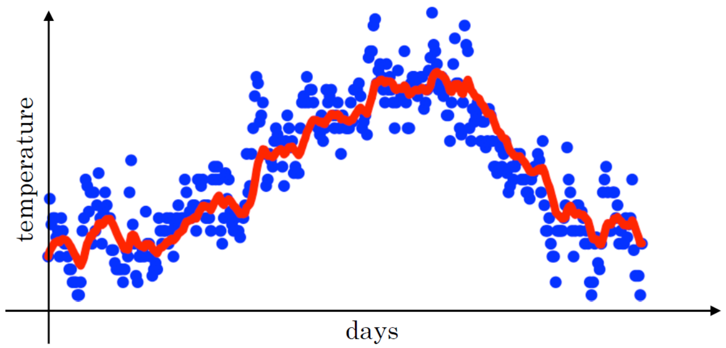
Exponentially Weighted Averages

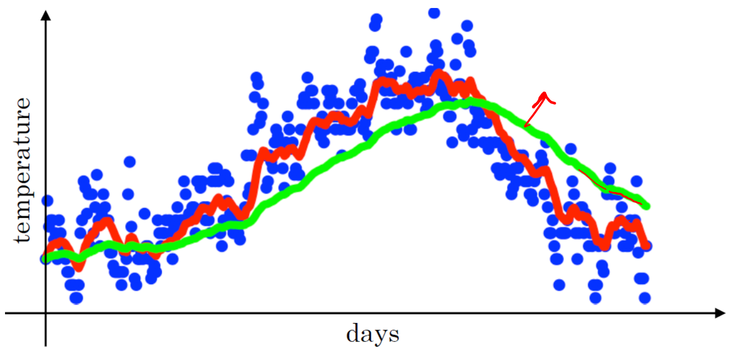
$$\boxed{v_t = \beta v_{t-1} + (1 - \beta)\theta_t}$$

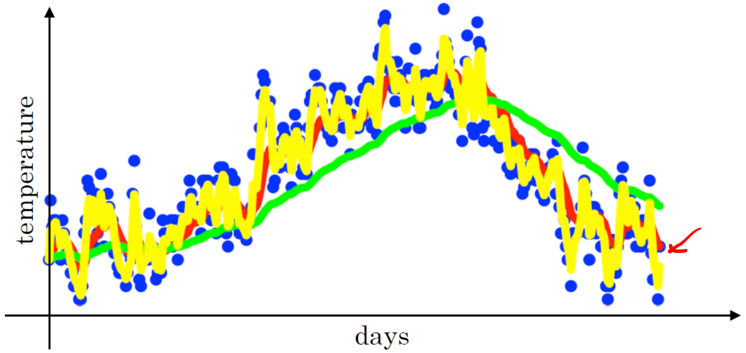
$\beta = 0.9$

$\beta = 0.98$

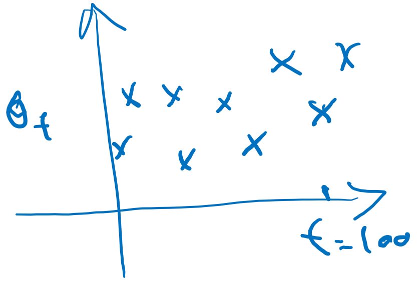
$\beta = 0.5$





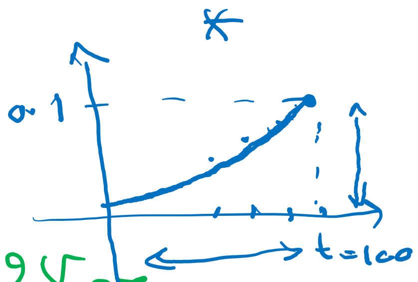


Exponentially Weighted Averages



$\varepsilon = 1-\beta$

$\approx \frac{1}{1-\beta}$



$$v_t = \beta v_{t-1} + (1-\beta)\theta_t$$

$$v_{100} = 0.9v_{99} + 0.1\theta_{100}$$
$$v_{99} = 0.9v_{98} + 0.1\theta_{99}$$
$$v_{98} = 0.9v_{97} + 0.1\theta_{98}$$

…

$\beta = 0.9 \rightarrow 10$

$\beta = 0.98 \rightarrow 50$

$0.1\,\theta_{98} + 0.9\,v_{97}$

$v_{100} = 0.1\,\theta_{100} + 0.9\,\cancel{v_{99}}\;(0.1\,\theta_{99} + 0.9\,\cancel{v_{98}})$

$= 0.1\,\theta_{100} + 0.1 \times 0.9\,\theta_{99} + 0.1\,(0.9)^2\,\theta_{98} + 0.1\,(0.9)^3\,\theta_{97} + 0.1\,(0.9)^4\,\theta_{96} \ldots$

$0.9^{10} \simeq 0.35 \simeq \frac{1}{e}$

$(1-\varepsilon)^{1/\varepsilon} = 1/e$

$\varepsilon = 0.1$

$(1-0.1)^{1/0.1} = 1/e$, $1/0.1 = 10$, $1-0.1 = 0.9$

$\beta = 0.98$

$\varepsilon = 0.02$

$0.98^{50} \simeq \frac{1}{e}$

Implementing Exponentially Weighted Averages

$$v_0 = 0$$
$$v_1 = \beta v_0 + (1-\beta)\,\theta_1$$
$$v_2 = \beta v_1 + (1-\beta)\,\theta_2$$
$$v_3 = \beta v_2 + (1-\beta)\,\theta_3$$
$$\ldots$$

$V_\theta = 0$

$V_\theta = \beta V_\theta + (1-\beta)\,\theta_1$

$V_\theta = \beta V_\theta + (1-\beta)\,\theta_2$

$\vdots$

$V_\theta = 0$

Keep repeat {

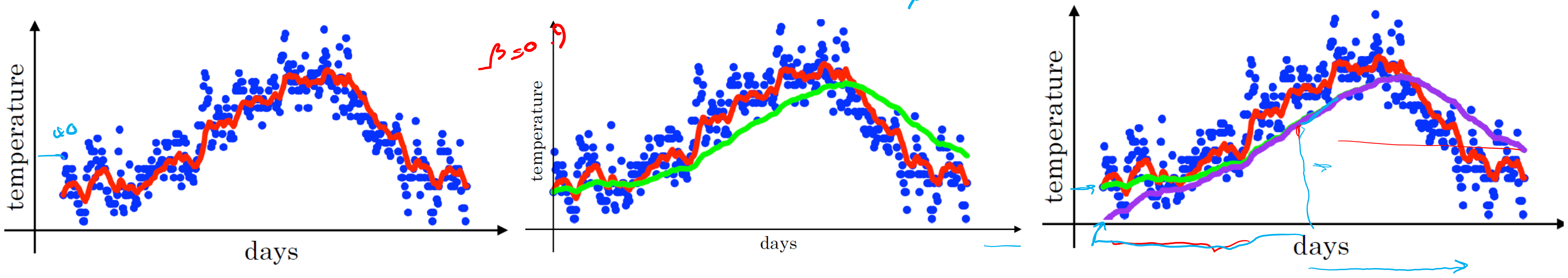
Get next θ_t

$V_\theta = \beta V_\theta + (1-\beta)\,\theta_t$

}

Optimization Algorithms: Bias Correction in Exponentially Weighted Averages

Bias Correction



$$v_t = \beta v_{t-1} + (1-\beta)\theta_t$$

$v_0 = 0$

$v_1 = \cancel{0.98\, v_0} + 0.02\,\theta_1$

$v_2 = 0.98\, v_1 + 0.02\,\theta_2$

$= 0.98 \times 0.02 \times \theta_1 + 0.02\,\theta_2$

$v_t = \dfrac{v_t}{1-\beta^t}$

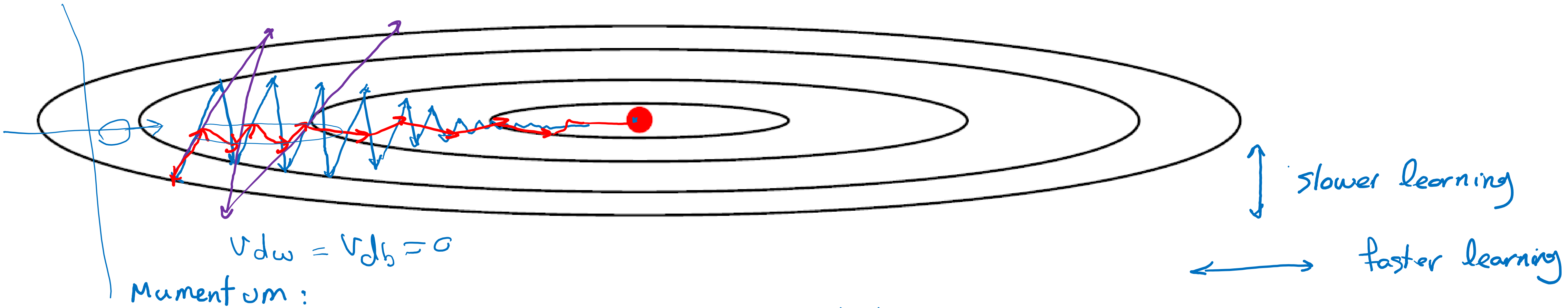
$t = 2$ $\quad 1-\beta^t = 1-(0.98)^2 = 0.039$

$v_2 = \dfrac{v_2}{0.039} = \dfrac{0.0196\,\theta_1 + 0.02\,\theta_2}{0.0396}$

$\beta < 1 \Rightarrow \beta^{t \to \infty} \to 0$

Optimization Algorithms: Gradient Descent with Momentum

Gradient Descent Example



$V_{dw} = V_{db} = 0$

Mumentum :

on iteration t :

compute dw, db on ~~batch~~ mini_batch

$V_{dw} = \beta V_{dw} + (1-\beta) dw$

$V_{db} = \beta V_{db} + (1-\beta) db$

$w := w - \alpha V_{dw}$, $b = b - \alpha V_{db}$

Implementation Details

$v_{db} = V_{dw} = 0$

On iteration t:

Compute dW, db on the current mini-batch

$$\boxed{v_{dW} = \beta v_{dW} + 1 - \beta \ dW}$$

$$v_{db} = \beta v_{db} + 1 - \beta \ dh$$

$$W = W - \alpha v_{dW}, \ b = b - \alpha v_{db}$$

$V_{dw} = \beta V_{dw} + d_w$

$\cancel{\dfrac{V_{dw}}{1-\beta^t}}$

$\dfrac{\alpha}{1-\beta}$

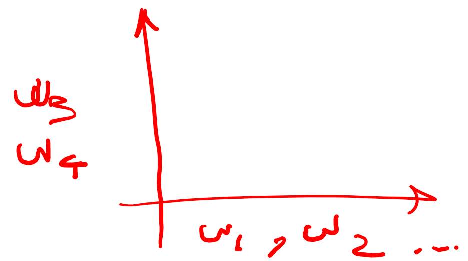
Hyperparameters: α, β

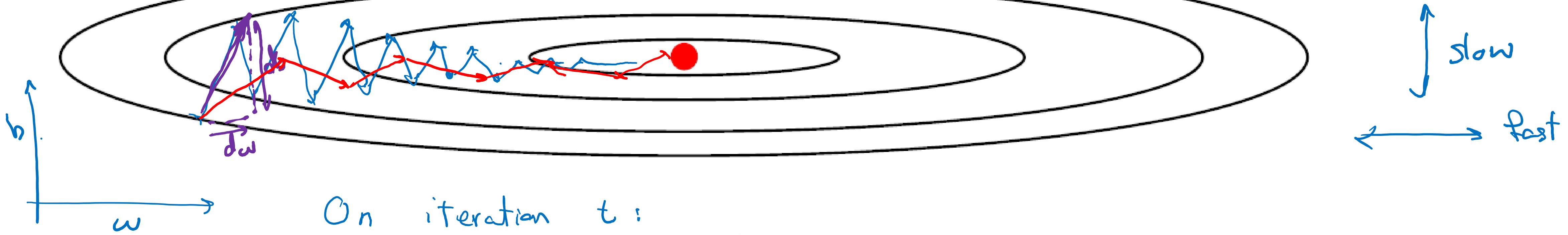
$\beta = 0.9$

average over ≈ 10 last iteration

Optimization Algorithms: RMSprop

RMSprop





On iteration t:

Compute dw, db on current mini-batch

$$\rightarrow S_{dw} = \beta_2 S_{dw} + (1-\beta_2)\, dw^2$$ → element-wise → small

$$S_{db} = \beta_2 S_{db} + (1-\beta_2)\, db^2$$ ← large

$$w := w - \alpha \frac{dw}{\sqrt{S_{dw}}+\varepsilon}$$ ← larger

$$b := b - \alpha \frac{db}{\sqrt{S_{db}}+\varepsilon}$$ → smaller

$\varepsilon = 10^{-8}$

Core Foundation Review

- Optimization Algorithms
  - Mini-batch Gradient Descent
  - Understanding Mini-batch Gradient Descent
  - Exponentially Weighted Averages
  - Understanding Exponentially Weighted Averages
  - Bias Correction in Exponentially Weighted Averages
  - Gradient Descent with Momentum
  - RMSprop