# Neural network training

For a given dataset $\mathcal{D}_n$, we aim to minimize the empirical risk

$$\hat{R}(\theta) = \hat{R}(f_\theta) = \frac{1}{n}\sum_{j=1}^{n} \ell(f_\theta(x_j), y_j) \qquad \text{(training error)}$$

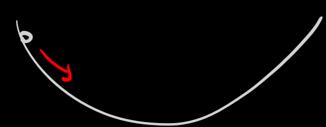
For an algorithm $\mathcal{A}$, we aim to characterize the optimization error

$$\hat{R}(\mathcal{A}(\mathcal{D}_n)) - \inf_{f \in \mathcal{F}} \hat{R}(f)$$

While in some cases this can be done explicitly (e.g. linear regression) in general the problem of finding the minimum of a function $\hat{R}$ is hard.

One method to (hopefully) solve this problem: move sequentially in the direction (in $\Theta$) of steepest descent of $\theta$ by updating

$$\theta \leftarrow \theta - \gamma D_\theta R(\theta)$$



for a small timestep parameter $\gamma$. This method is called gradient descent:

the update reads:

$$\theta_{k+1} = \theta_k - \gamma_k D_\theta R(\theta_k). \qquad \text{(GD)}$$

<u>Note</u>: Why using this and not trying to solve $D_\theta R(\theta) = 0$?

Computation of $D_\theta R$ is cheap: consider

$$D_{W^1_{11}} f(\theta) = \sigma'_{L+1}(z_{L+1}(x))\, W^L \sigma'_L(z_L(x)) \dots \sigma'_1(z_1(x)) \cdot x$$

In the above update, provided that we know $\sigma'_\ell$ and $z_\ell$ we are computing a complicated derivative by taking a product of known numbers ($z_\ell$ were evaluated to find $f_\theta(x)$) so that

$$\vartheta_{k+1} = \vartheta_k - \gamma_k D_\vartheta \hat{R}(\vartheta_k)$$

$$= \vartheta_k - \gamma_k \frac{1}{n} \sum_{j=1}^{n} \underbrace{D_\vartheta f_{\vartheta_k}(x_j)}_{\text{"easy"}} \underbrace{D_f \ell(f_{\vartheta_k}(x_j), y_j)}_{\text{"easy": just evaluate } f_{\vartheta_k}}$$

To study the dynamics of $\vartheta_k$ we write formally

$$\vartheta_{k+1} = \vartheta_k - \gamma D_\vartheta R(\vartheta_k) \longrightarrow \frac{\vartheta_{k+1} - \vartheta_k}{\gamma} = - D_\vartheta R(\vartheta_k)$$

Based on the above, one expects that $\vartheta_k \approx \bar{\vartheta}_{k\gamma}$ where $\bar{\vartheta} : \mathbb{R}_+ \to \Theta$ solves

$$\begin{cases} \dfrac{d}{dt} \bar{\vartheta}_t = - D_\vartheta R(\bar{\vartheta}_t) \\ \bar{\vartheta}_0 = \vartheta_0 \end{cases} \qquad \text{(gradient flow)}$$

<u>Def</u> a differentiable function $F : \Theta \to \mathbb{R}$ is L-Lipschitz smooth if

$$\| F(\vartheta') - F(\vartheta) - D_\vartheta F(\vartheta)(\vartheta' - \vartheta) \| \leq \frac{L}{2} \| \vartheta - \vartheta' \|^2 \quad \text{for a } L \geq 0$$

<u>Lemma</u>: F L-Lipschitz smooth $\Longleftrightarrow$ $D_\vartheta F$ is L-Lipschitz continuous

<u>Pf</u>: "$\Leftarrow$" Taylor $F(\vartheta') = F(\vartheta) + \int_\vartheta^{\vartheta'} D_\vartheta F(\eta) \, d$

$$= F(\vartheta) + \int_\vartheta^{\vartheta'} D_\vartheta F(\vartheta) + (D_\vartheta F(\eta) - D_\vartheta F(\vartheta))$$

$$= F(\vartheta) + D_\vartheta F(\vartheta)(\vartheta' - \vartheta) + \int_\vartheta^{\vartheta'} D_\vartheta F(\eta) - D_\vartheta F(\vartheta) d$$

$$\leq F(\vartheta) + D_\vartheta F(\vartheta)(\vartheta' - \vartheta) + \frac{L}{2} \|\vartheta - \vartheta'\|^2$$

$$\text{"}\Longrightarrow\text{"} \quad (D_\vartheta F(\vartheta) - D_\vartheta F(\vartheta'))(\vartheta - \vartheta')$$

Remark: Let $F \in C^2(\Theta)$. Then if $F$ is L-Lipschitz smooth we have

$$-L\mathbb{1} \preceq \| D^2_\vartheta F(\vartheta) \| \preceq L\mathbb{1}$$

Indeed for $\|\eta\| = 1$ : $\| \eta^T D^2_\vartheta F(\vartheta) \| = \lim_{h \to 0} \left\| \frac{D_\vartheta F(\vartheta + h\eta) - D_\vartheta F(\vartheta)}{h} \right\| \leq \frac{L\|h\eta\|}{h} = L$

Lemma: The solution to (GF) exists and is unique

Pf: Consequence of classical existence and uniqueness of solution to ODEs with Lipschitz vector fields.

From now on we set $F = R$

Lemma: Let $D_\vartheta R$ be Lipschitz. For every $T > 0$ there exists $C > 0$ s.t

$$\| \bar{\vartheta}_{k\gamma} - \vartheta_k \| \leq C\gamma \qquad \forall k \in \{0, \dots, \lfloor \tfrac{T}{\gamma} \rfloor\}$$

Proof: $$\bar{\vartheta}_{t+\gamma} = \bar{\vartheta}_t - \gamma D_\vartheta R(\bar{\vartheta}_t) + \underbrace{\int_t^{t+\gamma} D_\vartheta R(\bar{\vartheta}_s) - D_\vartheta R(\bar{\vartheta}_t)\, ds}_{\gamma^2}$$

And use that

$$\left\| \int_t^{t+\gamma} D_\vartheta R(\bar{\vartheta}_s) - D_\vartheta R(\bar{\vartheta}_t)\, ds \right\| \leq \gamma \sup_{s \in (t, t+\gamma)} \| D_\vartheta R(\bar{\vartheta}_s) - D_\vartheta R(\bar{\vartheta}_t) \| \leq \gamma \sup_{s \in (t, t+\gamma)} \| \bar{\vartheta}_s - \bar{\vartheta}_t \|$$

$$\le \gamma^2 \sup_{s\in[t,t+\gamma)} \|D_\theta R(\bar\theta_s)\|$$

By Lipschitz smoothness we have that $\sup_{t\in[0,T]} D_\theta R(\bar\theta_t) \le C$

$$e_{k+1} = \|\bar\theta_{(k+1)\gamma} - \theta_{k+1}\| = \|\bar\theta_{k\gamma} - \gamma D_\theta R(\bar\theta_{k\gamma}) - \theta_k + \gamma D_\theta R(\theta_k)\| + C\gamma^2$$

$$\le (1+\gamma\lambda)e_k + C\gamma^2$$

$$\Longrightarrow e_{k+1} = C\gamma^2 \sum_{j=1}^{k}(1+\lambda\gamma)^j = \frac{C\gamma^2}{\lambda\gamma}\left((1+\lambda\gamma)^{k+1}-1\right) \le \frac{C\gamma}{\lambda}\left(e^{\lambda\gamma(k+1)}-1\right) \le C'\gamma \qquad \square$$

This result justifies the use of GF instead of GD.

Q: when does (GF) converge, and when can we bound

$$\hat R(f_\theta) - \inf_{f\in\mathcal F}\hat R(f) \qquad ?$$

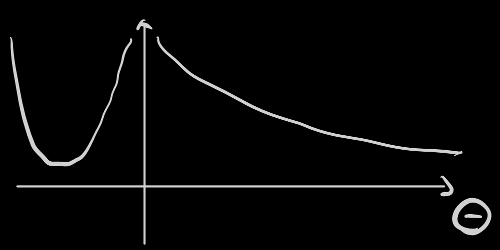
<u>Remark</u>: Since

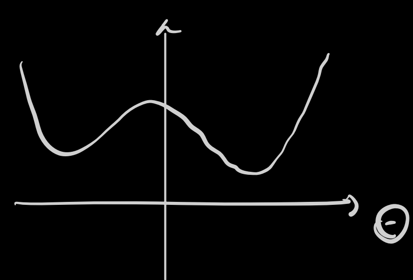
$$\frac{d}{dt}R(\theta_t) = D_\theta R(\theta_t)\cdot(-D_\theta R(\theta_t)) = -\|D_\theta R(\theta_t)\| \le 0$$

we have $R(\theta_t) \le R(\theta_0)$.

So if $R$ is bounded from below, as $t\to\infty$ by monotone convergence theorem $R(\theta_t)$ converges.

⚠ this does NOT show that $\theta_t$ converges!





However, we can't say much about optimality of limit points.

Example (linear regression) $n \geq d$ $\quad X^T X \succeq \lambda \mathbb{1}$

$$\hat{R}(\vartheta) = \frac{1}{n} \| X\vartheta - y \|_2^2 \Longrightarrow \vartheta_* = (X^T X)^{-1} X^T y$$

$$\longrightarrow R(\vartheta_*) = \frac{1}{n} \| (X(X^T X)^{-1} X - \mathbb{1}) y \|_2^2$$

$$\frac{d}{dt} \| \vartheta_t - \vartheta_* \|_2^2 = \langle \vartheta_t - \vartheta_*, -D_\vartheta \hat{R}(\vartheta_t) \rangle = -\langle \vartheta_t - \vartheta_*, X^T (X\vartheta_t - y) \rangle$$

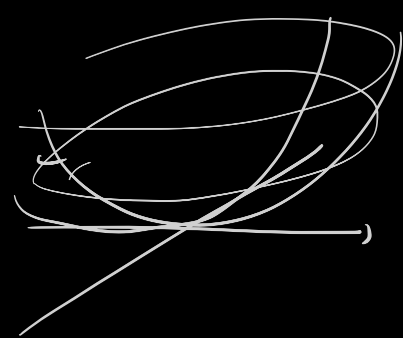
$$= -\langle \vartheta_t - \vartheta_*, X^T X (\vartheta_t - \vartheta_*) \rangle \leq -\lambda \| \vartheta_t - \vartheta_* \|_2^2$$

$$\xrightarrow{G} \| \vartheta_t - \vartheta_* \|_2^2 \leq e^{-\lambda t} \| \vartheta_0 - \vartheta_* \|_2^2$$

This happens because $\hat{R}$ is strongly convex:

$$\hat{R}(\vartheta) - \hat{R}(\vartheta_*) = D_\vartheta R(\vartheta_*) \cdot (\vartheta - \vartheta_*) + \frac{1}{2} (\vartheta - \vartheta_*) D_\vartheta^2 R(\vartheta_*) (\vartheta - \vartheta_*)$$

$$= \frac{1}{2} (\vartheta - \vartheta_*) X^T X (\vartheta - \vartheta_*).$$



The convergence properties of GF are guaranteed when $R$ is convex:

Def: a differentiable $\hat{R}: \Theta \to \mathbb{R}$ is $\lambda$-strongly convex for $\lambda \in \mathbb{R}$ if

$$(*) \qquad \hat{R}(\vartheta') \geq \hat{R}(\vartheta) + \langle D_\vartheta \hat{R}(\vartheta), \vartheta' - \vartheta \rangle + \frac{\lambda}{2}\|\vartheta' - \vartheta\|^2 \qquad \text{for any } \vartheta, \vartheta' \in \Theta$$

Lemma: If $R$ is convex ($\lambda$-strongly convex with $\lambda = 0$) if $DR(\vartheta_*) = 0$, $\vartheta_* \in \min R$

Pf: trivial from definition.

Lemma Let $\hat{R}$ be $\lambda$-convex, then the following holds:

$$\langle D_\vartheta \hat{R}(\vartheta') - D_\vartheta \hat{R}(\vartheta), \vartheta' - \vartheta \rangle \geq \lambda \|\vartheta - \vartheta'\|^2$$

Proof: Consider:

$$\hat{R}(\vartheta') \geq \hat{R}(\vartheta) + \langle D_\vartheta \hat{R}(\vartheta), \vartheta - \vartheta' \rangle + \frac{\lambda}{2}\|\vartheta - \vartheta'\|^2$$

adding the above to the definition of $\lambda$ strongly convex we are done $\square$

Thm: Let $R$ be $\lambda$-strongly convex and $\vartheta_*$ be the minimizer, then

$$\|D_\vartheta R(\vartheta)\|_2^2 \geq 2\lambda (R(\vartheta) - R(\vartheta_*))$$

Proof: Setting $\vartheta' = \vartheta - \frac{1}{\lambda} D_\vartheta R(\vartheta)$ in $(*)$ we get

$$R(\vartheta_*) \geq R(\vartheta) - \frac{1}{\lambda}\|D_\vartheta R(\vartheta)\|^2 + \frac{1}{2\lambda}\|D_\vartheta R(\vartheta)\|^2 = R(\vartheta) - \frac{1}{2\lambda}\|D_\vartheta R(\vartheta)\|^2$$

Thm: Let $R(\vartheta)$ be $\lambda$-convex, L-smooth, then there exists a unique $\vartheta^*$ and

$$\|\bar{\vartheta}_t - \vartheta_*\|^2 \leq \|\bar{\vartheta}_0 - \vartheta_*\|^2 e^{-2\lambda t}$$

$$R(\bar{\vartheta}_t) - R(\vartheta_*) \leq (R(\bar{\vartheta}_0) - R(\vartheta_*))e^{-2\lambda t}$$

Proof: $\frac{d}{dt}\frac{1}{2}\|\bar\vartheta_t - \vartheta_*\|^2 = \langle \bar\vartheta_t - \vartheta_*, \frac{d}{dt}\vartheta_t \rangle$

$$= -\langle \bar\vartheta_t - \vartheta_*, D_\vartheta R(\bar\vartheta_t) - D_\vartheta R(\vartheta_*) \rangle$$

$$\leq -\lambda \|\bar\vartheta_t - \vartheta_*\|^2$$

$\longrightarrow$ Grönwall

Similarly $\frac{d}{dt} R(\vartheta_t) - R(\vartheta_*) = \langle D_\vartheta R(\vartheta_t), \dot\vartheta_t \rangle$

$$= -\|D_\vartheta R(\vartheta_t)\|^2$$

Thm $\longrightarrow$ $\leq -2\lambda (R(\vartheta_t) - R(\vartheta_*))$

$\longrightarrow$ Grönwall

Lemma: Assume $R$ is $L$-smooth, $\lambda$ strongly convex.

Choose $\gamma = \frac{1}{L}$. Then $R(\vartheta_t) - R(\vartheta_*) \leq e^{-\frac{t}{\kappa}}(\ldots)$

Pf: $R(\vartheta_t) = R(\vartheta_{t-1} - \frac{1}{L} D R(\vartheta_{t-1})) \leq R(\vartheta_{t-1}) - \frac{1}{L}\|D_\vartheta R(\vartheta_{t-1})\|^2$

$$+ \frac{1}{2L}\|D_\vartheta R(\vartheta_{t-1})\|^2$$

$\Longrightarrow R(\vartheta_t) - R(\vartheta_*) \leq R(\vartheta_{t-1}) - R(\vartheta_*) - \frac{1}{2L}\|D_\vartheta R(\vartheta_{t-1})\|^2$

$$\leq \left(1 - \frac{\lambda}{2L}\right)(R(\vartheta_{t-1}) - R(\vartheta_*))$$

$\square$

# Gradient flows for convex functionals.

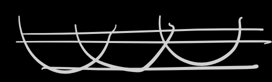
**Example**: Consider now the case where some eigenvalues of $X^TX$ are $0$.

$$\frac{d}{dt}\vartheta_t - \vartheta_* = -X^TX(\vartheta_t - \vartheta_*) \Rightarrow \vartheta_t - \vartheta_* = e^{-X^TXt}(\vartheta_0 - \vartheta_*)$$

$$R(\vartheta_t) - R(\vartheta_*) = \left(\frac{1}{2}(\vartheta_t - \vartheta_*)X^TX(\vartheta_t - \vartheta_*)\right) = \frac{1}{2}(\vartheta_0 - \vartheta_*)\underbrace{e^{-X^TXt}X^TXe^{-X^TXt}}_{A(t)}(\vartheta_0 - \vartheta_*)$$

$$\mathrm{Eig}(A(t)) = e^{-\lambda t}\cdot\lambda\cdot e^{-\lambda t} = \frac{1}{t}(\lambda t)e^{-\lambda t} \leq \frac{1}{t}$$

**Theorem**: Let $R(\vartheta)$ be convex and Lipschitz smooth. Then



$$R(\bar{\vartheta}_t) - R(\bar{\vartheta}_*) \leq C\frac{1}{t}.$$

**Proof**: $R(\vartheta') \geq R(\vartheta) - D_\vartheta R(\vartheta)(\vartheta - \vartheta')$

$$\Rightarrow \frac{d}{dt}\frac{1}{2}\|\vartheta_t - \vartheta_*\|^2 = (\vartheta_t - \vartheta_*)(-D_\vartheta R(\vartheta_t))$$

$$\leq R(\vartheta^*) - R(\vartheta_t)$$

$$\Rightarrow \int_0^t R(\vartheta_s) - R(\vartheta_*)\,ds \leq -\frac{1}{2}\|\vartheta_t - \vartheta_*\| + \frac{1}{2}\|\vartheta_0 - \vartheta_*\| \leq \frac{1}{2}\|\vartheta_0 - \vartheta_*\|$$

$$\Rightarrow \frac{1}{t}\int_0^t R(\vartheta_s)\,ds - R(\vartheta_*) \leq \frac{1}{2t}C_0$$

but $\frac{1}{t}\int_0^t R(\vartheta_s)\,ds \geq R(\vartheta_t)$ $\qquad$ $\frac{d}{dt}R(\vartheta_t) \leq 0$

$$\Rightarrow R(\vartheta_t) - R(\vartheta_*) \leq \frac{1}{2t}C_0$$

$\square$

Lemma: Let $\bar{\theta}_t$ converge to $\theta_*$ and there exists $N(\theta_*)$ s.t. $D_\theta R(\theta) \succeq \varepsilon^2 \mathbb{1}$ $\forall \theta \in N$ then for $\gamma$ sufficiently small

$$\sup_t \| \bar{\theta}_t - \theta_t \| \leq C\gamma$$

Pf: Let $T$ be such that $\bar{\theta}_t \in N$.