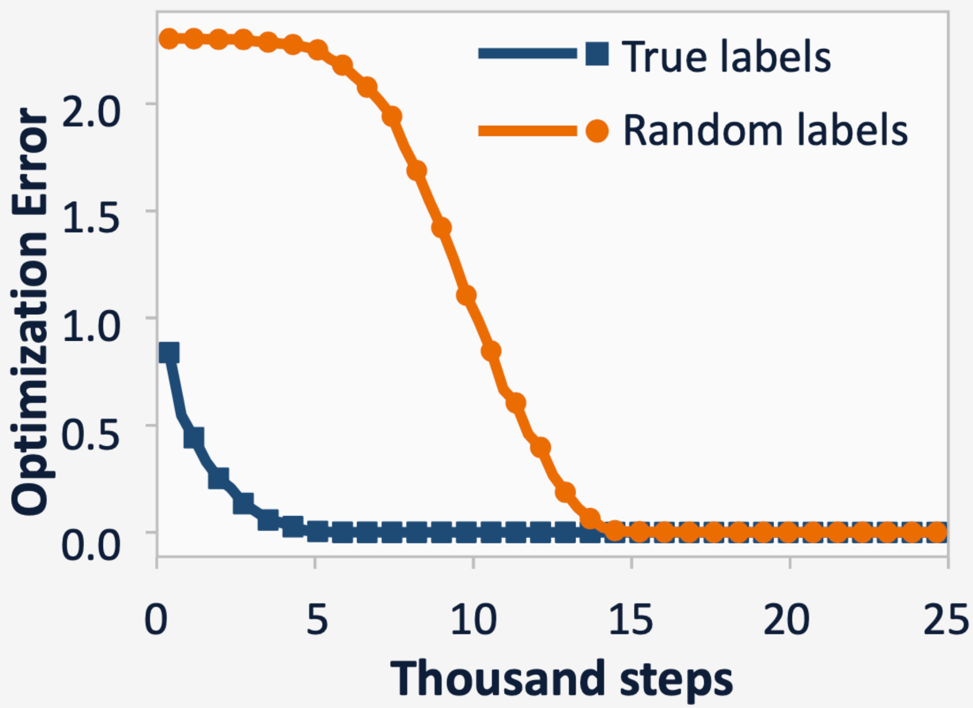
# Non-convex Optimization Landscape

# Gradient descent finds global minima

**Practice:** gradient descent

$$\theta(t+1) \leftarrow \theta(t) - \eta \frac{\partial L(\theta(t))}{\partial \theta(t)}$$

# parameter >> n



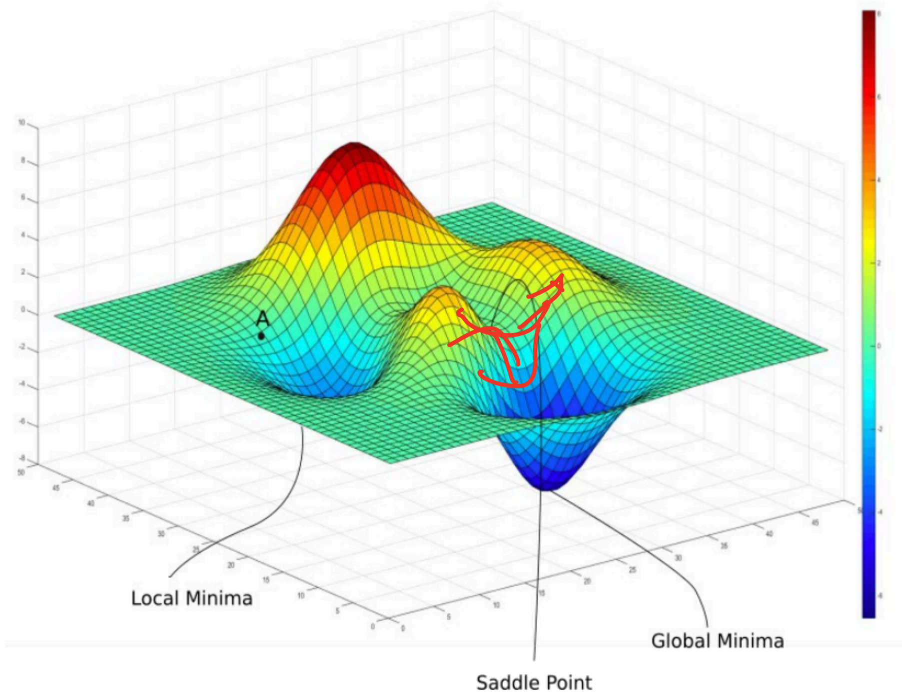

Optimization error **-> 0** for both ***true labels*** and ***random labels*** !

Zhang Bengio Hardt Recht Vinyals 2017

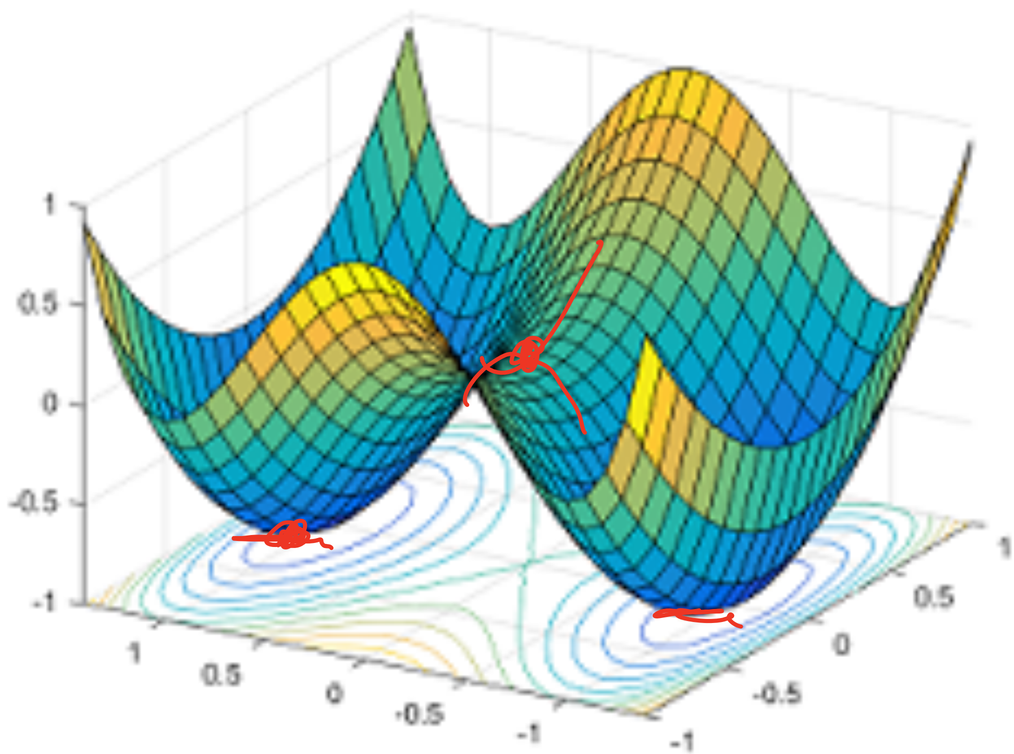
Understanding DL Requires Rethinking Generalization

# Types of stationary points

- Stationary points: $x : \nabla f(x) = 0$
- Global minimum:
  $x : f(x) \leq f(x') \forall x' \in \mathbb{R}^d$
- Local minimum:
  $x : f(x) \leq f(x') \forall x' : \|x - x'\| \leq \epsilon$
- Local maximum:
  $x : f(x) \geq f(x') \forall x' : \|x - x'\| \leq \epsilon$
- Saddle points: stationary points that are not a local min/max



# Landscape Analysis

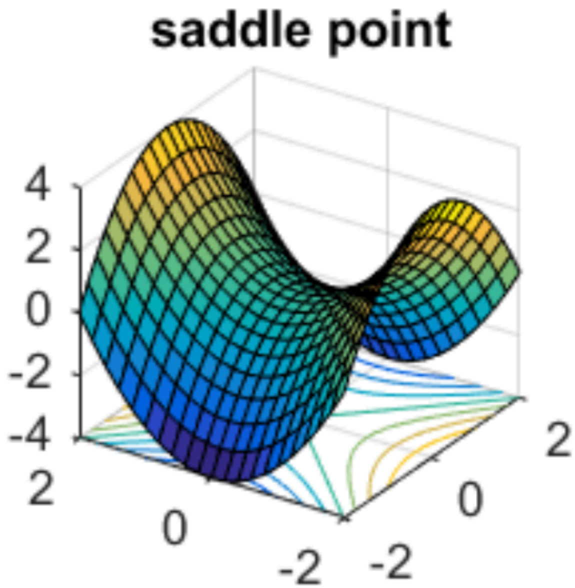



hope

- All local minima are global!
- Gradient descent can escape saddle points.

# Strict Saddle Points (Ge et al. ’15, Sun et al. ’15)

$v^T A$
$= \lambda_{min} \cdot v$




$\nabla f(x) = 0$

- Strict saddle point: a saddle point and $\lambda_{\min}(\nabla^2 f(x)) < 0$

min $f(x) = \frac{1}{2} x^T A x$, $\lambda_{min}(A) < 0$, eigenvector $v$: unit vector

$x = 0$, $\nabla f(x) = Ax = 0$, $\nabla^2 f(x) = A$

$x_0$ close to $0$, $x_{t+1} = x_t - \eta \nabla f(x_t)$ suppose $\eta$ sufficiently small

$\|x_{t+1}\|_2 \geq |v^T x_{t+1}|$

$= |v^T (x_t - \eta A x_t)|$

$= |v^T x_t - \eta \cdot \lambda_{min} v^T x_t| = \underbrace{|1 - \eta \lambda_{min}(A)|}_{>1} \cdot |v^T x_t|$

exp ↑

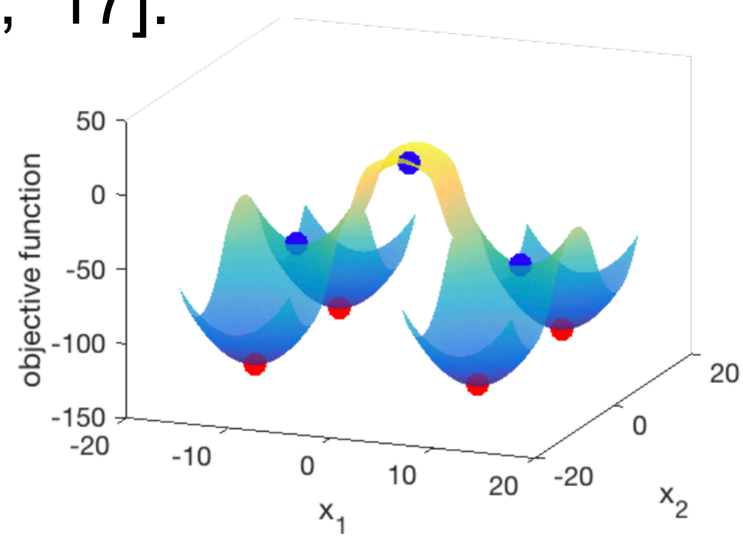
# Escaping Strict Saddle Points

$x_{t+1} = x_t - \eta \nabla f(x_t) + \eta \cdot \varepsilon$

$\varepsilon \sim \mathcal{N}(0, I)$

- Noise-injected gradient descent can escape strict saddle points in polynomial time [Ge et al., '15, Jin et al., '17].
- Randomly initialized gradient descent can escape all strict saddle points asymptotically [Lee et al., '15].
  - Stable manifold theorem. $x_0 \sim$ randomly initialized
- Randomly initialized gradient descent can take exponential time to escape strict saddle points [Du et al., '17].

If 1) all local minima are global, and 2) are saddle points are strict, then noise-injected (stochastic) gradient descent finds a global minimum in polynomial time



# What problems satisfy these two conditions

- Matrix factorization

$$\min_{U,V} \|UV^T - A\|_F^2$$

- Matrix sensing

$$\sum_{i=1}^{n} \left( \langle S_i, UU^T \rangle - y_i \right)^2$$

- Matrix completion

$$\begin{bmatrix} * & * \\ * & * \end{bmatrix}$$ low-rank

- Tensor factorization

$$v^{\otimes 3}$$

- Two-layer neural network with quadratic activation

$$f(x_i) = \sum_{j=1}^{m} \langle x_i, w_j \rangle^2$$

# What about neural networks?

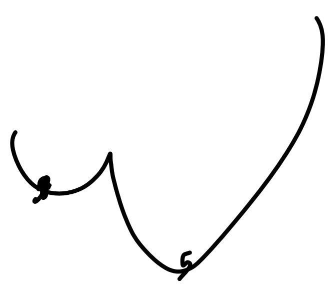
$x \in \mathbb{R}^d$, $d$ large

- Linear networks (neural networks with linear activations functions): all local minima are global, but there exists saddle points that are not strict [Kawaguchi '16].

$\exists x,\ \lambda_{min}(\nabla^2 f(x)) = 0$ $\quad \min_{W_1, \dots, W_{l-1}} (W_l W_{l-1} \cdots W_1 x - y)^2, \quad \min_{w} (wx - y)^2$

- Non-linear neural networks with:
  - Virtually any non-linearity, ReLU, sigmoid
  - Even with Gaussian inputs, $x \sim N(0, I)$
  - Labels are generated by a neural network of the same architecture, $y = NN(x)$

There are many bad local minima [Safran-Shamir '18, Yun-Sra-Jadbaie '19].



# Global convergence of gradient descent

# Global convergence of gradient descent

convex loss

**Theorem** (Du et al. ’18, Allen-Zhu et al. ’18, Zou et al ’19) If the width of each layer is poly(n) where n is the number of data. Using random initialization with a particular scaling, gradient descent finds an approximate global minimum in polynomial time.

$\varepsilon$ - global min

poly(n) $\log(\frac{1}{\varepsilon})$ for quadratic

neural tangent kernel

proof for a two-layer NN

# Gradient Flow: a Kernel Point of View

- $L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(\theta, x_i), y_i)$

$$\frac{\partial L(\theta)}{\partial \theta} = \frac{1}{n} \sum_{i=1}^{n} \ell'(f(\theta, x_i), y_i) \cdot \frac{\partial f(\theta, x_i)}{\partial \theta}$$

GF: $\frac{d\theta(t)}{dt} = -\frac{\partial L(\theta)}{\partial \theta}$

if $L(\theta)$ strongly convex, $\exists$ unique $\theta^*$, $\theta(t) \to \theta^*$

for NN, # of parameter, $\dim(\theta) > n$

we want show, $t \to \infty$, $f(\theta(t), x_i) \to y_i$

# Gradient Flow: a Kernel Point of View

$u_i(t) = f(\theta(t), x_i)$, $\quad u(t) = \begin{pmatrix} u_1(t) \\ \vdots \\ u_n(t) \end{pmatrix}$

$$\frac{du_i(t)}{dt} = \left\langle \frac{\partial u_i(t)}{\partial \theta_t}, \frac{d\theta(t)}{dt} \right\rangle$$

$l'(u(t), y) \in \mathbb{R}^n$

$[l'(u(t), y)]_i$

$= l'(u_i(t), y_i)$

$H(t) \in \mathbb{R}^{n \times n}$

$$= \left\langle \frac{\partial u_i(t)}{\partial \theta(t)}, -\frac{1}{n} \sum_{j=1}^{n} l'(u_j(t), y_j) \cdot \frac{\partial u_j(t)}{\partial \theta(t)} \right\rangle$$

$$= -\frac{1}{n} [l'(u_1(t), y_1), \ldots, l'(u_n(t), y_n)] \cdot \left( \left\langle \frac{\partial u_i(t)}{\partial \theta(t)}, \frac{\partial u_1(t)}{\partial \theta(t)} \right\rangle, \ldots, \left\langle \frac{\partial u_i(t)}{\partial \theta(t)}, \frac{\partial u_n(t)}{\partial \theta(t)} \right\rangle \right)$$

$[H(t)]_{ij}$

$$= \left\langle \frac{\partial u_i(t)}{\partial \theta(t)}, \frac{\partial u_j(t)}{\partial \theta(t)} \right\rangle,$$

$$\boxed{\frac{du(t)}{dt} = -\frac{1}{n} H(t) \cdot l'(u(t), y)}$$

# Gradient Flow: a Kernel Point of View

If $l$ is quadratic, $l(u(t), y) = \frac{1}{2}(u(t) - y)^2$

$$l'(u(t), y) = u(t) - y$$

$$\frac{d(u(t) - y)}{dt} = -\frac{1}{n} H(t)(u(t) - y)$$

If $H(t)$ is always positive definite

$\forall t$, $\lambda_{min}(H(t)) \geq \lambda_0$, $\lambda_0 > 0$

$$\rightarrow \frac{1}{2}\|u(t) - y\|_2^2 \rightarrow 0$$

Pf: $$\frac{d\left(\frac{1}{2}\|u(t) - y\|_2^2\right)}{dt} = -\frac{1}{n}(u(t) - y)^T H(t)(u(t) - y)$$

$$\leq -\frac{\lambda_0}{n}\|u(t) - y\|_2^2$$

$H$ p.d.

$\lambda_{min}(H) \geq \lambda_0$

any vector $u$

$u^T H u \geq \lambda_0 \|u\|_2^2$

# Gradient Flow: a Kernel Point of View

Consider $\frac{d}{dt}\left(\exp\left(\frac{\lambda_0 t}{n}\right)\cdot\frac{1}{2}\|u(t)-y\|_2^2\right)$

$= \frac{\lambda_0}{2n}\exp\left(\frac{\lambda_0 t}{n}\right)\|u(t)-y\|_2^2 + \frac{d\left(\frac{1}{2}\|u(t)-y\|_2^2\right)}{dt}\exp\left(\frac{\lambda_0 t}{n}\right)$

$\leq \exp\left(\frac{\lambda_0 t}{n}\right)\|u(t)-y\|_2^2\left(\frac{\lambda_0 t}{2n}-\frac{\lambda_0 t}{n}\right) < 0$

$\Rightarrow \exp\left(\frac{\lambda_0 t}{n}\right)\cdot\frac{1}{2}\|u(t)-y\|_2^2$ is decreasing

$t=0,\quad \frac{1}{2}\|u(0)-y\|_2^2 \quad O(1)$

$\forall t \quad \exp\left(\frac{\lambda_0 t}{n}\right)\cdot\frac{1}{2}\|u(t)-y\|_2^2 \leq C$

$\Rightarrow \quad \frac{1}{2}\|u(t)-y\|_2^2 \leq C\cdot\exp\left(-\frac{\lambda_0 t}{n}\right)$

$\log\left(\frac{1}{\varepsilon}\right) \quad t\to\infty,\ \text{loss}\to 0,\ u(t)-y$

# Gradient Flow: a Kernel Point of View

$$f(\Theta, x) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \cdot \sigma(w_r^T x),$$

$m$: width, $x \in \mathbb{R}^d$, $a_r \in \mathbb{R}$, $w_r \in \mathbb{R}^d$, $\sigma(\cdot)$: ReLU

- Initialization: $a_r \sim \mathrm{unif}\{1, -1\}$ for simplicity

  $w_r \sim N(0, I)$

- Training: only train $w_1, \ldots, w_m$

$$\min_{w_1, \ldots, w_m} \frac{1}{n} \sum_{i=1}^{n} \left( f(x_i, a, w) - y_i \right)^2$$

$$u_i(t) = f(x_i, a, w(t))$$

$$\frac{du(t)}{dt} = -\frac{1}{n} H(t) (u(t) - y)$$

$H^*$: NTK

Idea: $H(t)$ stays the same for $\forall t$

$H(t) \approx H^*$

$$H^*_{ij} = \lim_{m \to \infty} \mathbb{E}_{\substack{\text{init} \\ \text{dist}}} \left\langle \frac{\partial f(x_i)}{\partial \theta}, \frac{\partial f(x_j)}{\partial \theta} \right\rangle$$

# Gradient Flow: a Kernel Point of View

$$H_{ij}(t) = \left\langle \frac{\partial u_i(t)}{\partial W(t)}, \frac{\partial u_j(t)}{\partial W(t)} \right\rangle, \quad W \in \mathbb{R}^{m \times d}$$

$$= \sum_{r=1}^{m} \left\langle \frac{\partial u_i(t)}{\partial w_r(t)}, \frac{\partial u_j(t)}{\partial w_r(t)} \right\rangle$$

$$\frac{\partial u_i(t)}{\partial w_r(t)} = \frac{1}{\sqrt{m}} a_r \cdot x_i \cdot \mathbb{1}\{w_r^T x_i \geq 0\}$$

$$H_{ij}(t) = \sum_{r=1}^{m} \frac{1}{m} \langle a_r x_i \mathbb{1}\{w_r^T x_i \geq 0\}, a_r x_j \mathbb{1}\{w_r^T x_j \geq 0\} \rangle$$

$$= \frac{1}{m} x_i^T x_j \sum_{r=1}^{m} \mathbb{1}\{w_r^T x_i \geq 0, w_r^T x_j \geq 0\}$$

To show: $H(t) \approx H^*$, ① $H(0) \approx H^*$

② $H(t) \approx H(0), \forall t$