

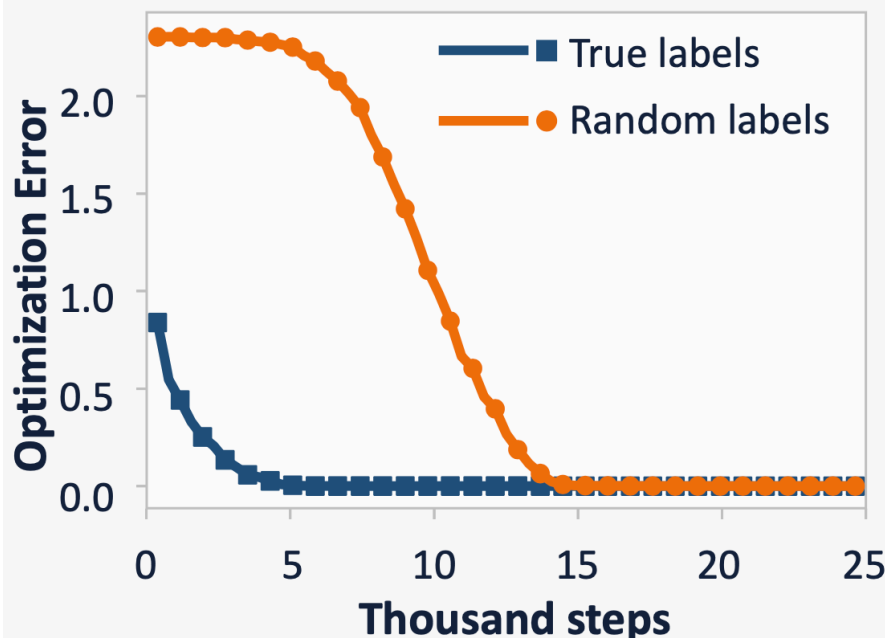
Non-convex Optimization Landscape

W

Gradient descent finds global minima

Practice: gradient descent

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



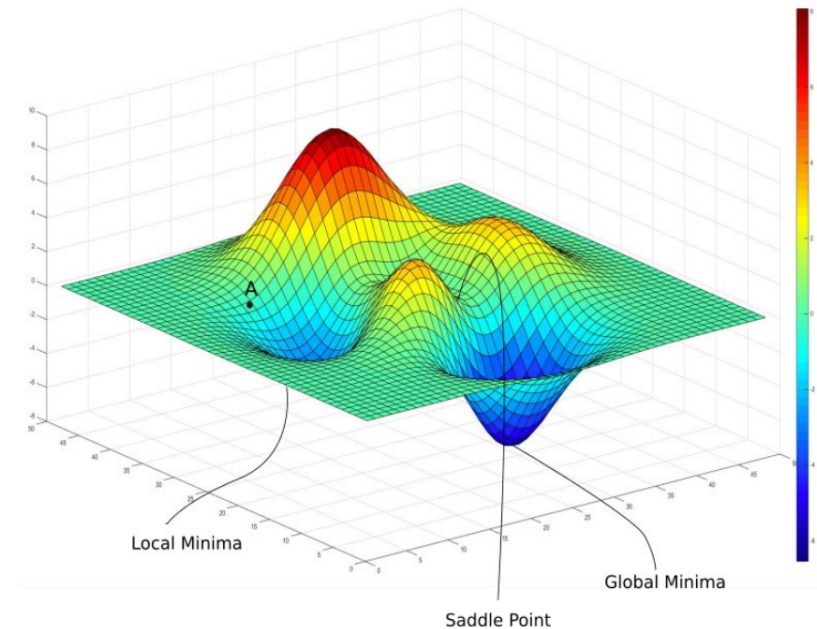
Optimization error $\rightarrow 0$ for both **true labels** and **random labels** !

Zhang Bengio Hardt Recht Vinyals 2017

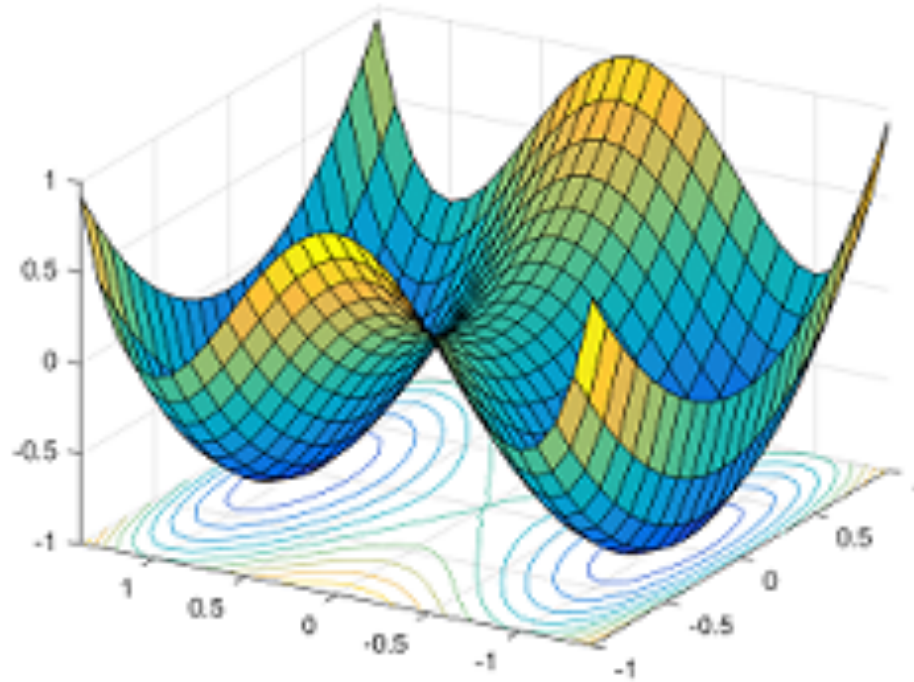
Understanding DL Requires Rethinking Generalization

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

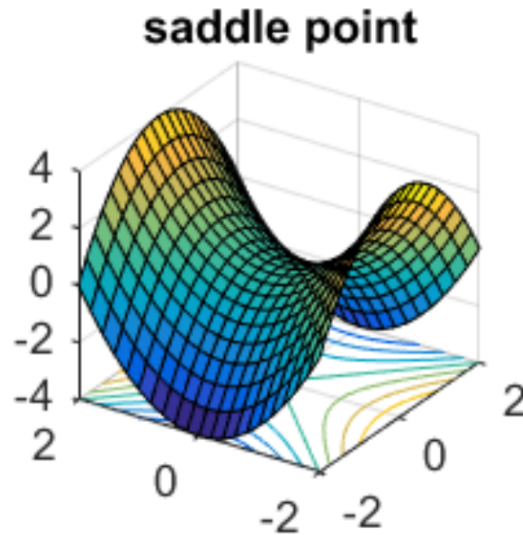


Landscape Analysis



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

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

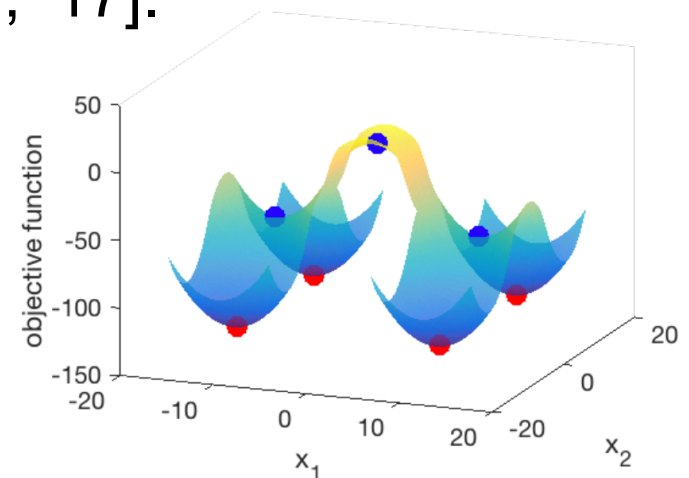


- Strict saddle point: a saddle point and $\lambda_{\min}(\nabla^2 f(x)) < 0$

Escaping Strict Saddle Points

- **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.
- 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
- Matrix sensing
- Matrix completion
- Tensor factorization
- Two-layer neural network with quadratic activation

What about neural networks?

- Linear networks (neural networks with linear activations functions): **all local minima are global, but there exists saddle points that are not strict** [Kawaguchi '16].
 - Non-linear neural networks with:
 - Virtually any non-linearity,
 - Even with Gaussian inputs,
 - Labels are generated by a neural network of the same architecture,
- There are many bad local minima** [Safran-Shamir '18, Yun-Sra-Jadbaie '19].

Global convergence of gradient descent



Global convergence of gradient descent

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

Gradient Flow: a Kernel Point of View

Gradient Flow: a Kernel Point of View

Gradient Flow: a Kernel Point of View

Gradient Flow: a Kernel Point of View

Gradient Flow: a Kernel Point of View

Gradient Flow: a Kernel Point of View
