Optimization in Machine Learning: From Convexity to Non-Convexity

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France

COLT, July 2025

Machine learning
Scientific context

- **Proliferation of digital data**
  - Personal data
  - Industry
  - Scientific: from bioinformatics to humanities

- **Need for automated processing of massive data, and beyond**

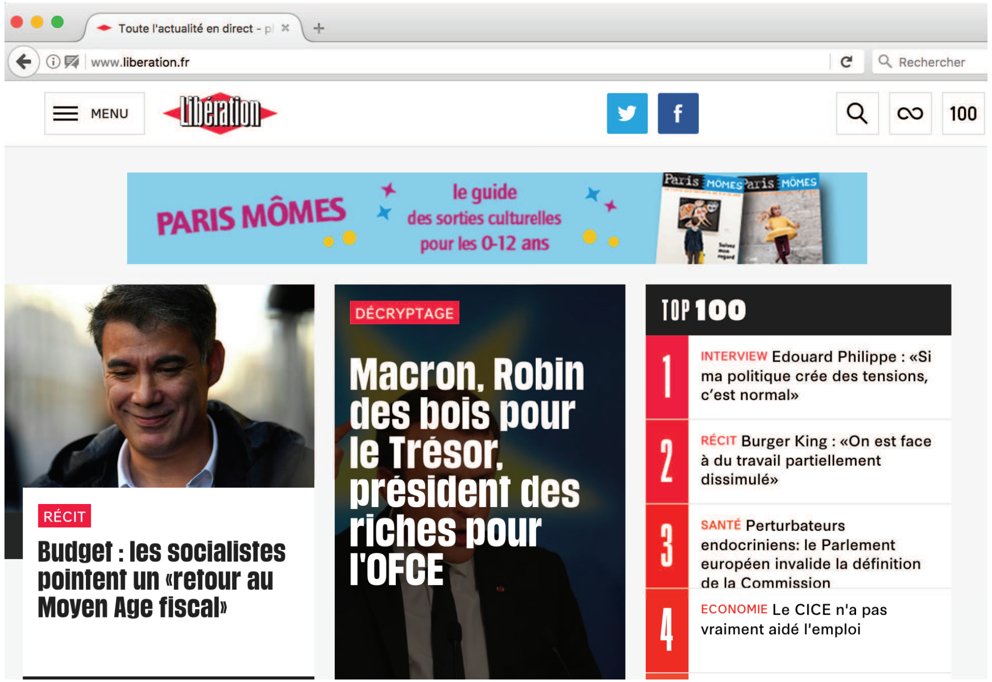
- **Series of "hypes"**

  Big data $\rightarrow$ Data science $\rightarrow$ Machine Learning
  $\rightarrow$ Deep Learning $\rightarrow$ Artificial Intelligence $\rightarrow$ Large Language Models

- **Positioning of learning theory?**

Parametric supervised machine learning

- **Data**: n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$
- **Prediction function** $h(x, \theta) \in \mathbb{R}$ parameterized by $\theta \in \mathbb{R}^d$



- Linear predictions
  - $h(x, \theta) = \theta^\top \Phi(x) = \sum_{i=1}^{d} \theta_i \Phi(x)_i$
- E.g., **advertising**: $n > 10^9$
  - $\Phi(x) \in \{0, 1\}^d$, $d > 10^9$
  - Navigation history + ad

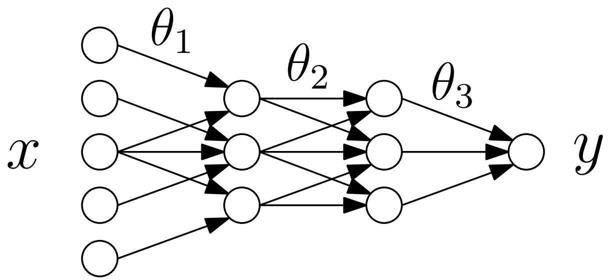
Parametric supervised machine learning

- **Data**: n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$
- **Prediction function** $h(x, \theta) \in \mathbb{R}$ parameterized by $\theta \in \mathbb{R}^d$

x_1 x_2 x_3 x_4 x_5 x_6

$y_1 = 1$ $y_2 = 1$ $y_3 = 1$ $y_4 = -1$ $y_5 = -1$ $y_6 = -1$

– Neural networks $(n, d > 10^8)$: $h(x, \theta) = \theta_r^\top \sigma(\theta_{r-1}^\top \sigma(\cdots \theta_2^\top \sigma(\theta_1^\top x))$



Parametric supervised machine learning

- **Data**: n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$
- **Prediction function** $h(x, \theta) \in \mathbb{R}$ parameterized by $\theta \in \mathbb{R}^d$
- **(regularized) empirical risk minimization**:

$$\min_{\theta \in \mathbb{R}^d} \quad \frac{1}{n} \sum_{i=1}^{n} \ell\big(y_i, h(x_i, \theta)\big) \quad + \quad \lambda \Omega(\theta)$$

data fitting term $\quad + \quad$ regularizer

Parametric supervised machine learning

- **Data**: n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \ldots, n$
- **Prediction function** $h(x, \theta) \in \mathbb{R}$ parameterized by $\theta \in \mathbb{R}^d$
- **(regularized) empirical risk minimization**:

$$\min_{\theta \in \mathbb{R}^d} \quad \frac{1}{n} \sum_{i=1}^{n} \big(y_i - h(x_i, \theta)\big)^2 \quad + \quad \lambda \Omega(\theta)$$

(least-squares regression)

Parametric supervised machine learning

- **Data**: n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$
- **Prediction function** $h(x, \theta) \in \mathbb{R}$ parameterized by $\theta \in \mathbb{R}^d$
- **(regularized) empirical risk minimization**:

$$\min_{\theta \in \mathbb{R}^d} \quad \frac{1}{n} \sum_{i=1}^{n} \log\big(1 + \exp(-y_i h(x_i, \theta))\big) \quad + \quad \lambda \Omega(\theta)$$

(logistic regression)

Parametric supervised machine learning

- **Data**: n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$, **independent, same distribution**
- **Prediction function** $h(x, \theta) \in \mathbb{R}$ parameterized by $\theta \in \mathbb{R}^d$
- **(regularized) empirical risk minimization**:

$$\min_{\theta \in \mathbb{R}^d} \quad \frac{1}{n} \sum_{i=1}^{n} \ell\big(y_i, h(x_i, \theta)\big) \quad + \quad \lambda \Omega(\theta)$$

data fitting term $+$ regularizer

- **Actual goal**: minimize test error $\mathbb{E}_{p(x,y)} \ell(y, h(x, \theta))$
  - Statistics and optimization

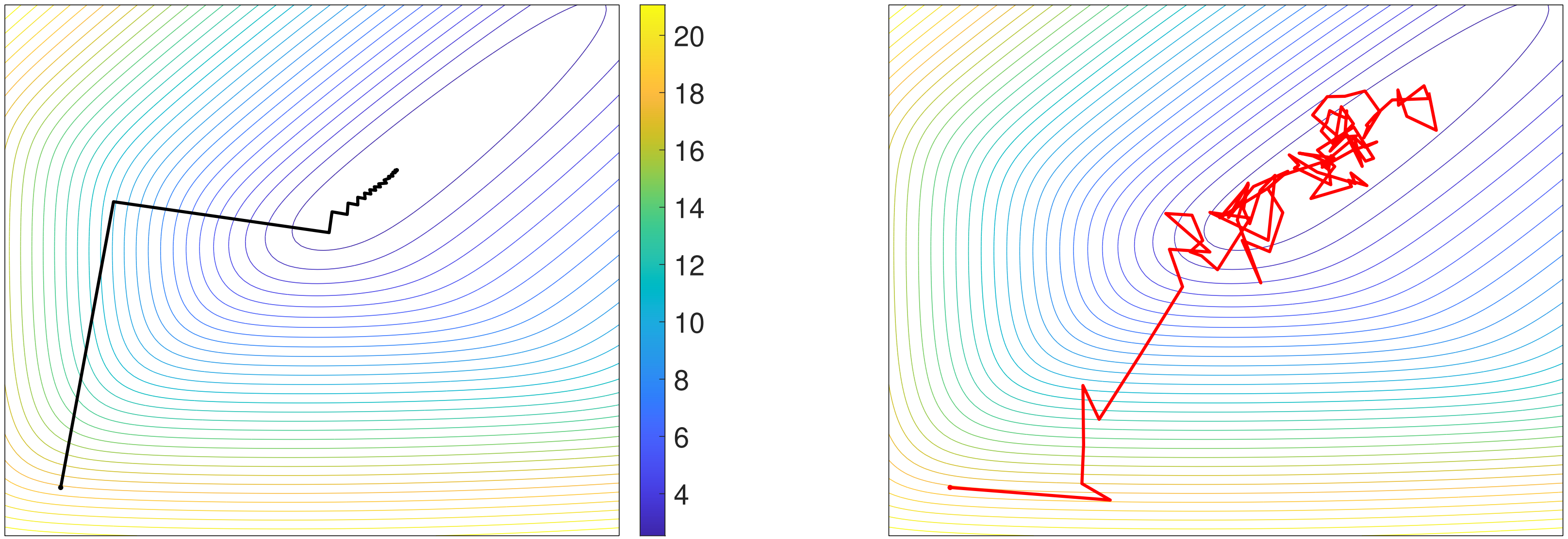
Convex optimization problems

$$\min_{\theta \in \mathbb{R}^d} \quad \frac{1}{n} \sum_{i=1}^{n} \ \ell\big(y_i, h(x_i, \theta)\big) \quad + \quad \lambda \Omega(\theta)$$

- **Conditions**: Convex loss (e.g., square) and "linear" predictions $h(x, \theta) = \theta^\top \Phi(x)$
- **Consequences**
  - Efficient algorithms (typically gradient-based)
  - Quantitative runtime and prediction performance guarantees
- **Golden years of convexity in machine learning** (1995 to 2020)
  - Support vector machines and kernel methods
  - Sparsity / low-rank models with first-order methods (Lasso, etc.)
  - Optimal transport
  - Stochastic methods for large-scale learning and online learning
  - etc.

Deterministic and stochastic methods

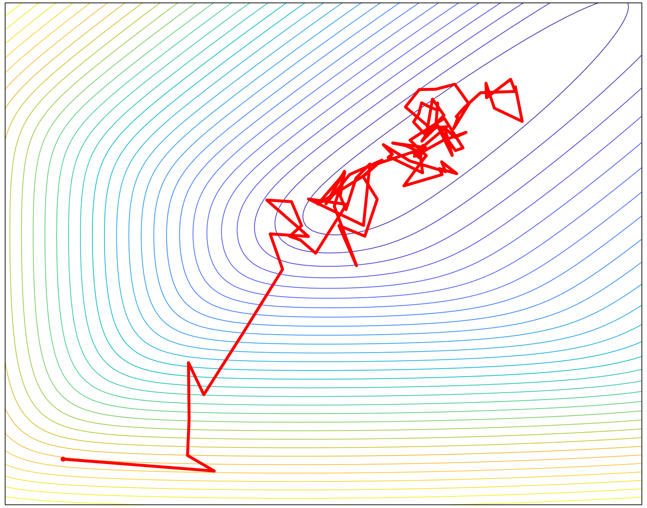
- Minimize $g(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$ with $f_i(\theta) = \ell\big(y_i, h(x_i,\theta)\big) + \lambda\Omega(\theta)$
- **Gradient descent**: $\theta_t = \theta_{t-1} - \gamma \nabla g(\theta_{t-1}) = \theta_{t-1} - \frac{\gamma}{n}\sum_{i=1}^{n} \nabla f_i(\theta_{t-1})$ (Cauchy, 1847)
- **Stochastic gradient descent**: $\theta_t = \theta_{t-1} - \gamma \nabla f_{i(t)}(\theta_{t-1})$ (Robbins and Monro, 1951)

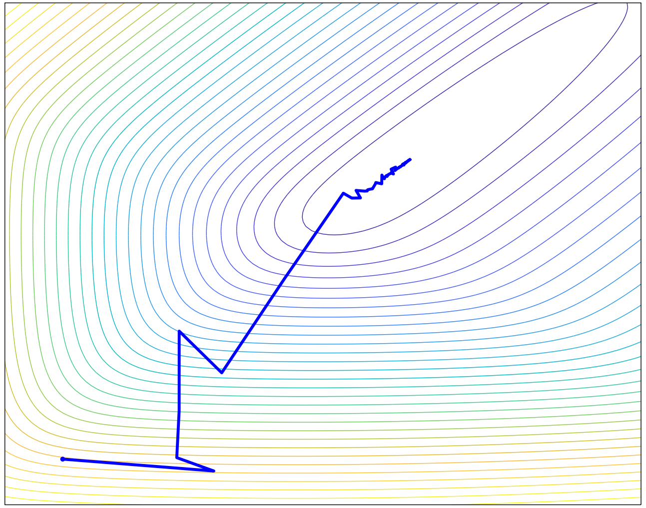


Stochastic gradient with exponential convergence

- **Variance reduction**
  - SAG (Le Roux, Schmidt, and Bach, 2012)
  - SVRG (Johnson and Zhang, 2013; Zhang, Mahdavi, and Jin, 2013)
  - SAGA (Defazio, Bach, and Lacoste-Julien, 2014)

$$\theta_t = \theta_{t-1} - \gamma \Big[\nabla f_{i(t)}(\theta_{t-1}) + \frac{1}{n} \sum_{i=1}^{n} y_i^{t-1} - y_{i(t)}^{t-1} \Big]$$





Stochastic gradient with exponential convergence

- **Variance reduction**
  - SAG (Le Roux, Schmidt, and Bach, 2012)
  - SVRG (Johnson and Zhang, 2013; Zhang, Mahdavi, and Jin, 2013)
  - SAGA (Defazio, Bach, and Lacoste-Julien, 2014)

- **Number of individual gradient computations to reach error ε** (strongly-convex objectives with condition number κ)

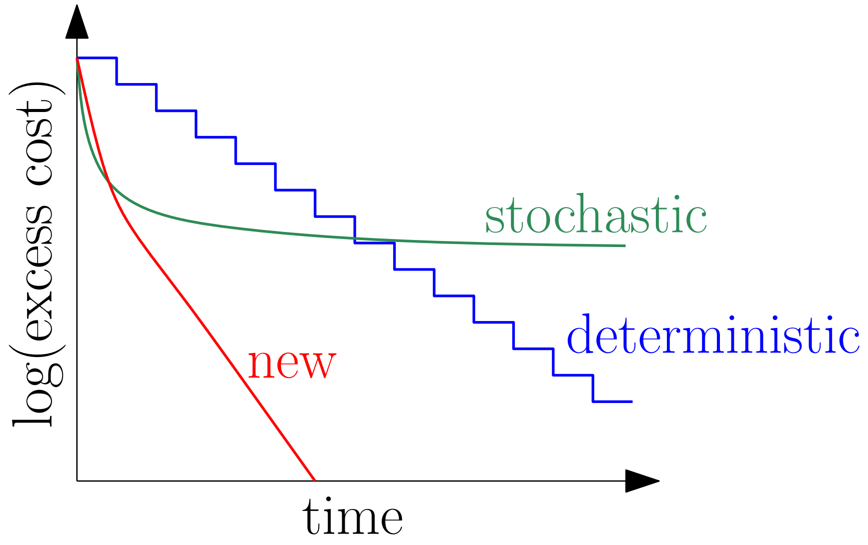
Gradient descent	$n\kappa \quad \times \log \frac{1}{\varepsilon}$
Stochastic gradient descent	$\kappa \quad \times \quad \frac{1}{\varepsilon}$
Variance reduction	$(n+\kappa) \quad \times \log \frac{1}{\varepsilon}$

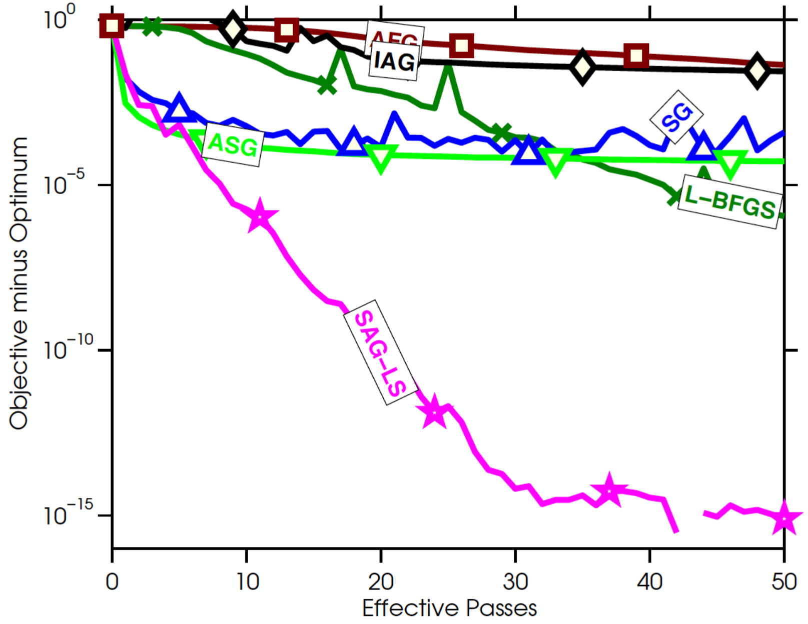
  - "Breaking" two lower bounds with extra assumptions

Stochastic gradient with exponential convergence

- **Acceleration** (Nesterov, 1983, 2004)
  - Shalev-Shwartz and Zhang (2014); Nitanda (2014); Lan (2015); Lin et al. (2015)
  - Optimal convergence rate: from $(n+\kappa)\cdot\log\frac{1}{\varepsilon}$ to $(n+\sqrt{n\kappa})\cdot\log\frac{1}{\varepsilon}$ gradient calls
- **Extension to online learning / single-pass SGD**
  - Nguyen et al. (2017); Fang et al. (2018); Cutkosky and Orabona (2019)
  - Guarantees beyond convex problems
- **Extensions to problems with finite sum structures**
  - Min-max saddle-point problems and variational inequalities (Balamurugan and Bach, 2016; Alacaoglu and Malitsky, 2022)
- **Extensions to distributed optimization** (e.g., Hendrikx, Bach, and Massoulié, 2019)

Stochastic gradient with exponential convergence
From theory to practice and vice-versa



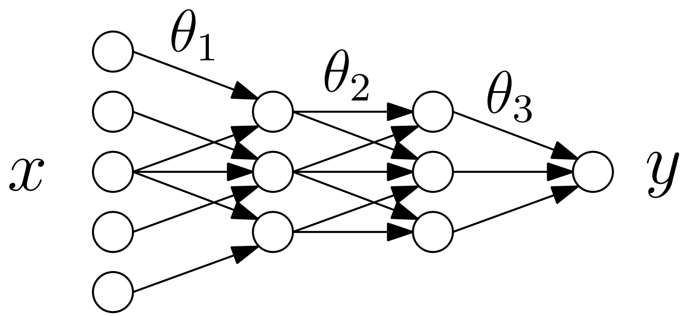


- **Empirical performance “matches” theoretical guarantees**
- **Theoretical analysis suggests practical improvements**
  - Non-uniform sampling, acceleration
  - Matching upper and lower bounds

What about deep learning?

Theoretical analysis of deep learning

- **Multi-layer neural network** $h(x,\theta) = \theta_r^\top \sigma(\theta_{r-1}^\top \sigma(\cdots \theta_2^\top \sigma(\theta_1^\top x))$

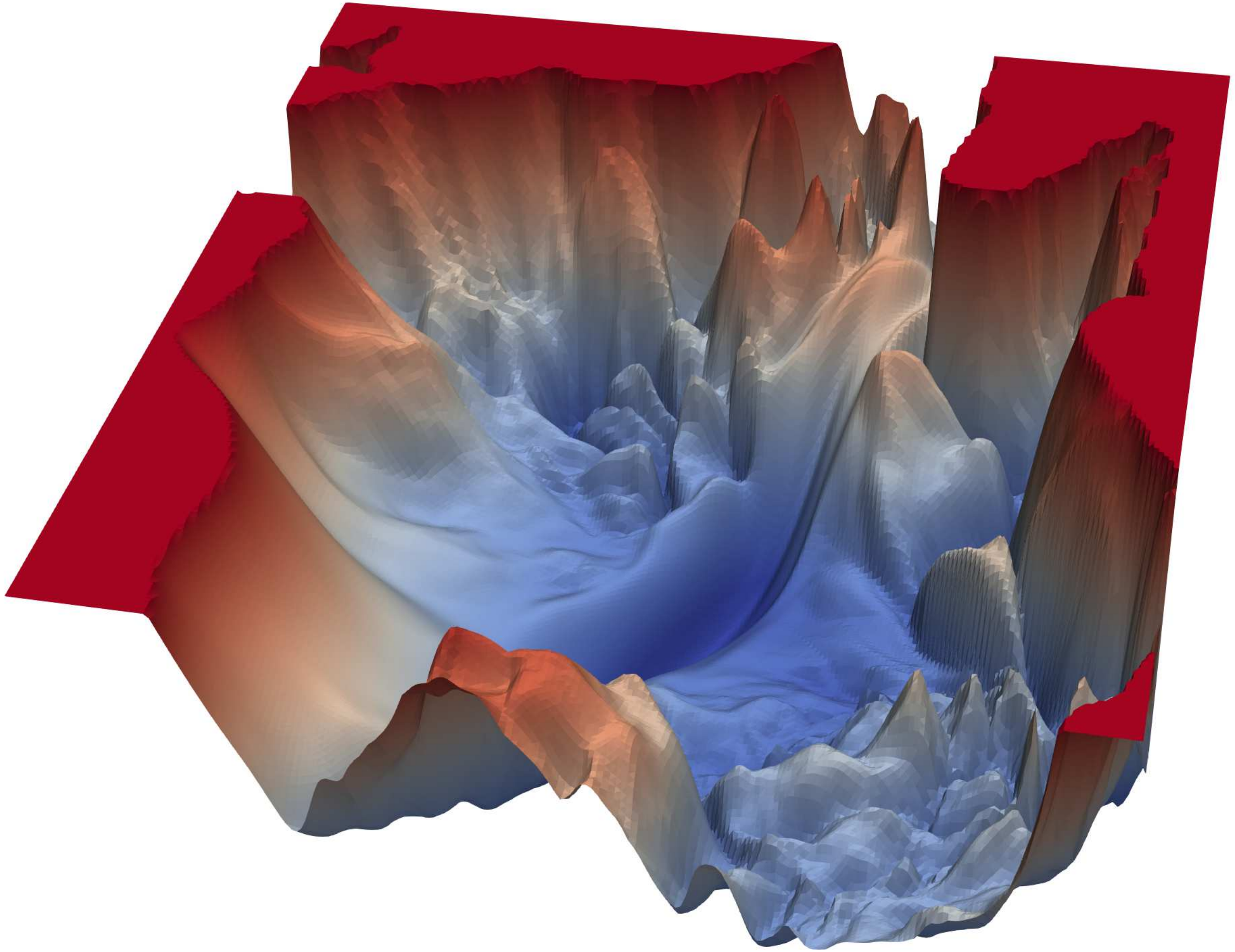


  – NB: already a simplification (see Resnets, transformers, Mamba, etc.)

- **Main difficulties**

  1. Non-convex optimization problems
  2. Generalization guarantees in the overparameterized regime

Loss landscape for deep learning (Li et al., 2018)



- **What can go wrong?**
  - Local minima
  - Stationary points
  - Plateaux
  - Bad initialization
  - etc...

Optimization algorithms for deep learning

$$\min_{\theta \in \mathbb{R}^d} g(\theta) = \frac{1}{n} \sum_{i=1}^{n} f_i(\theta) \quad \text{with} \quad f_i(\theta) = \ell\big(y_i, h(x_i, \theta)\big) + \lambda \Omega(\theta)$$

- **Stochastic gradient descent** (Robbins and Monro, 1951): $\theta_t = \theta_{t-1} - \gamma \nabla f_{i(t)}(\theta_{t-1})$
  - Mini-batches, momentum: $\theta_t = \theta_{t-1} - \gamma \nabla f_{i(t)}(\theta_{t-1}) + \delta(\theta_{t-1} - \theta_{t-2})$
  - Global guarantees in the convex case, local guarantees otherwise (see, e.g., Bottou et al., 2018)

- **Adam** (Kingma and Ba, 2014)
  - Rescaled updates with a reconditioning effect
  - Global guarantees in the convex case, local guarantees otherwise (Reddi et al., 2018; Défossez et al., 2020)

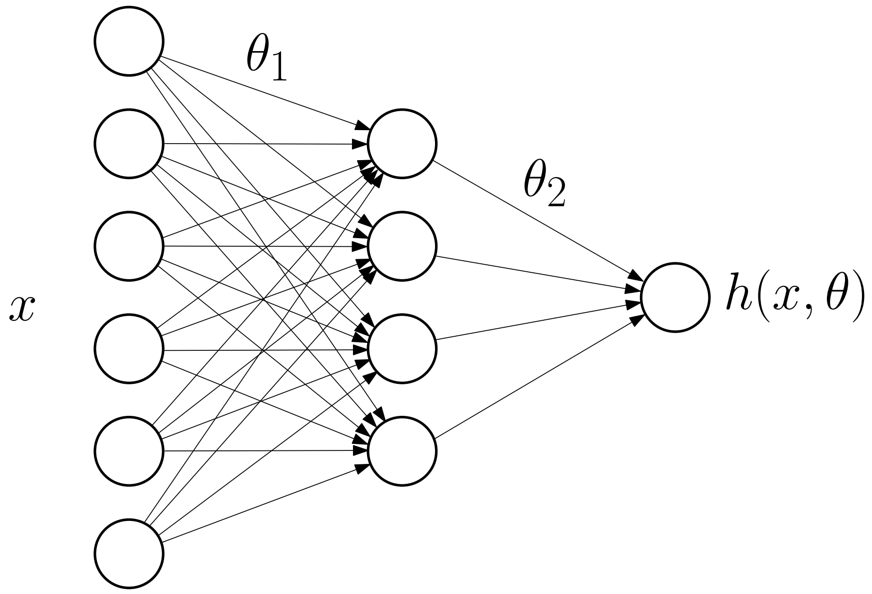
Optimization algorithms for deep learning

$$\min_{\theta \in \mathbb{R}^d} g(\theta) = \frac{1}{n} \sum_{i=1}^{n} f_i(\theta) \quad \text{with} \quad f_i(\theta) = \ell\big(y_i, h(x_i, \theta)\big) + \lambda \Omega(\theta)$$

- **Stochastic gradient descent** (Robbins and Monro, 1951): $\theta_t = \theta_{t-1} - \gamma \nabla f_{i(t)}(\theta_{t-1})$
  - Mini-batches, momentum: $\theta_t = \theta_{t-1} - \gamma \nabla f_{i(t)}(\theta_{t-1}) + \delta(\theta_{t-1} - \theta_{t-2})$
  - Global guarantees in the convex case, local guarantees otherwise (see, e.g., Bottou et al., 2018)
- **Adam** (Kingma and Ba, 2014)
  - Rescaled updates with a reconditioning effect
  - Global guarantees in the convex case, local guarantees otherwise (Reddi et al., 2018; Défossez et al., 2020)
- **Why does it work so well for overparameterized deep models?**

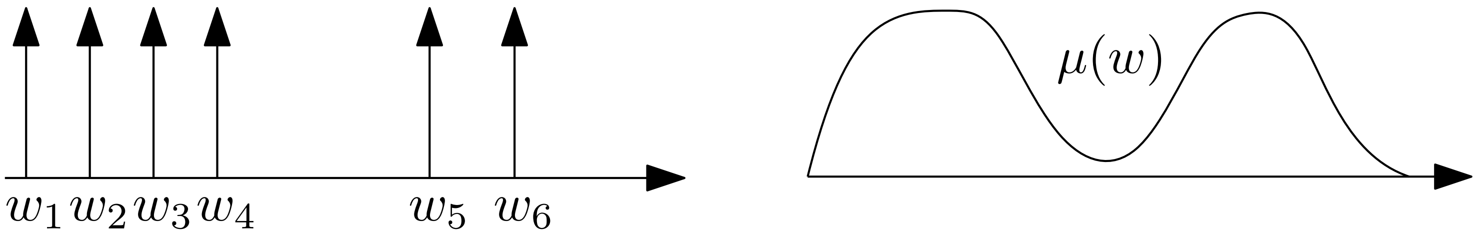
Gradient descent for a single hidden layer

- **Predictor**: $h(x) = \frac{1}{m}\theta_2^\top \sigma(\theta_1^\top x) = \frac{1}{m}\sum_{j=1}^m \theta_2(j) \cdot \sigma\big[\theta_1(\cdot, j)^\top x\big]$
  - Family: $h = \dfrac{1}{m}\displaystyle\sum_{j=1}^m \Psi(w_j)$ with $\Psi(w_j)(x) = \theta_2(j) \cdot \sigma\big[\theta_1(\cdot, j)^\top x\big]$
- **Goal**: minimize $R(h) = \mathbb{E}_{p(x,y)}\ell(y, h(x))$, with R convex



Gradient descent for a single hidden layer

- **Predictor**: $h(x) = \frac{1}{m}\theta_2^\top \sigma(\theta_1^\top x) = \frac{1}{m}\sum_{j=1}^m \theta_2(j) \cdot \sigma\big[\theta_1(\cdot, j)^\top x\big]$
  - Family: $h = \frac{1}{m}\sum_{j=1}^m \Psi(w_j)$ with $\Psi(w_j)(x) = \theta_2(j) \cdot \sigma\big[\theta_1(\cdot, j)^\top x\big]$
- **Goal**: minimize $R(h) = \mathbb{E}_{p(x,y)}\ell(y, h(x))$, with R convex
- **Main insight**
  - $h = \frac{1}{m}\sum_{j=1}^m \Psi(w_j) = \int_{\mathcal{W}} \Psi(w) d\mu(w)$ with $d\mu(w) = \frac{1}{m}\sum_{j=1}^m \delta_{w_j}$



Gradient descent for a single hidden layer

- **Predictor**: $h(x) = \frac{1}{m}\theta_2^\top \sigma(\theta_1^\top x) = \frac{1}{m}\sum_{j=1}^m \theta_2(j) \cdot \sigma\big[\theta_1(\cdot,j)^\top x\big]$
  - Family: $h = \displaystyle\frac{1}{m}\sum_{j=1}^m \Psi(w_j)$ with $\Psi(w_j)(x) = \theta_2(j) \cdot \sigma\big[\theta_1(\cdot,j)^\top x\big]$
- **Goal**: minimize $R(h) = \mathbb{E}_{p(x,y)}\ell(y,h(x))$, with R convex

- **Main insight**
  - $h = \displaystyle\frac{1}{m}\sum_{j=1}^m \Psi(w_j) = \int_{\mathcal{W}} \Psi(w) d\mu(w)$ with $d\mu(w) = \displaystyle\frac{1}{m}\sum_{j=1}^m \delta_{w_j}$
  - Overparameterized models with m large $\approx$ measure μ with densities
  - Barron (1993); Kurkova and Sanguineti (2001); Bengio et al. (2006); Rosset et al. (2007); Bach (2017)

Optimization on measures

- **Minimize with respect to measure** μ: $R\Big(\int_{\mathcal{W}} \Psi(w) d\mu(w)\Big)$
  - Convex optimization problem on measures
  - Frank-Wolfe techniques for incremental learning
  - Non-tractable (Bach, 2017), not what is used in practice

- **Represent** μ **by a finite set of "particles"** $\mu = \frac{1}{m}\sum_{j=1}^m \delta_{w_j}$
  - Backpropagation = gradient descent on $W = (w_1, \dots, w_m)$

- **Three questions**:
  - Algorithm limit when number of particles m gets large
  - Global convergence to a global minimizer
  - Prediction performance

Many particle limit and global convergence (Chizat and Bach, 2018)

- **General framework**: minimize $F(\mu) = R\Big(\int_{\mathcal{W}} \Psi(w) d\mu(w)\Big)$
  - Algorithm: minimizing $F_m(w_1, \dots, w_m) = R\Big(\frac{1}{m}\sum_{j=1}^{m} \Psi(w_j)\Big)$
  - Gradient flow $\dot{W} = -m \nabla F_m(W)$, with $W = (w_1, \dots, w_m)$
  - Idealization of (stochastic) gradient descent
    1. Single pass SGD on the unobserved expected risk
    2. Multiple pass SGD or full GD on the empirical risk

Many particle limit and global convergence (Chizat and Bach, 2018)

- **General framework**: minimize $F(\mu) = R\Big(\int_{\mathcal{W}} \Psi(w) d\mu(w)\Big)$
  - Algorithm: minimizing $F_m(w_1, \dots, w_m) = R\Big(\frac{1}{m}\sum_{j=1}^{m} \Psi(w_j)\Big)$
  - Gradient flow $\dot{W} = -m\nabla F_m(W)$, with $W = (w_1, \dots, w_m)$
  - Idealization of (stochastic) gradient descent
- **Limit when m tends to infinity**
  - Wasserstein gradient flow (Nitanda and Suzuki, 2017; Chizat and Bach, 2018; Mei, Montanari, and Nguyen, 2018; Sirignano and Spiliopoulos, 2018)

Many particle limit and global convergence (Chizat and Bach, 2018)

- **(informal) theorem**: when the number of hidden neurons tends to infinity, the gradient flow converges to the global optimum
  - One-hidden-layer neural networks and beyond
  - "Mean-field" limit common in statistical physics (Mei et al., 2018)
  - Two key ingredients: homogeneity and initialization, on top of convexity of the loss

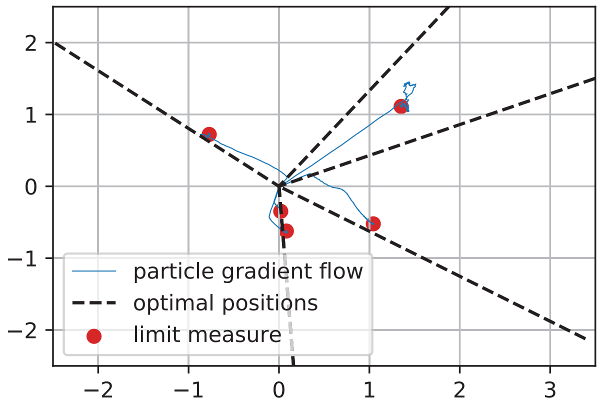
- **Homogeneity** (see, e.g., Haeffele and Vidal, 2017; Bach et al., 2008)
  - Rectified linear units: $\sigma(u) = \max\{u, 0\}$

- **Sufficiently diverse initial neuron weights**
  - Needs to cover the entire sphere of directions

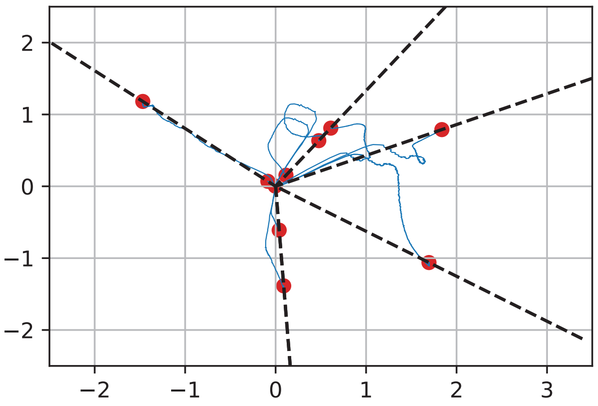
- **Blessing of overparameterization, but only qualititative**

Simple simulations in two dimensions

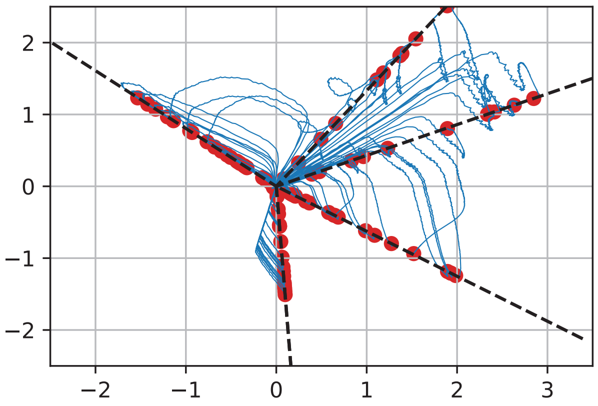
- ReLU units with $d = 2$ (optimal predictor has 5 neurons)



5 neurons



10 neurons



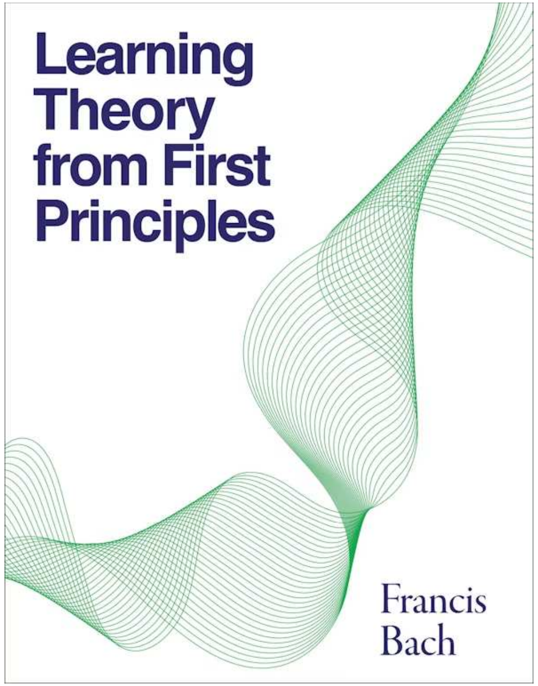
100 neurons

Model: $h(x, \theta) = \frac{1}{m} \sum_{j=1}^{m} \eta_j \max\{w_j^\top x, 0\}$

(plotting $|\eta_j| w_j$ for each hidden neuron j)

Avoiding overfitting with overparameterization

- **Common wisdom:** “models do not generalize well with too many parameters”
  - aggregated magnitude of parameters (e.g., a norm) provides a finer control
- **Regularization effect of gradient methods**
  - “Implicit bias” towards minimum norm solutions (Gunasekar et al., 2017; Soudry et al., 2018; Gunasekar et al., 2018; Ji and Telgarsky, 2018; Chizat and Bach, 2020)
  - No catastrophic, benign overfitting (Bartlett et al., 2020)



(MIT Press, 2024)

Towards quantitative convergence

- **Convergence time t and number m of neurons required for global convergence**
  - Too hard (yet!) in general
  - Complex dynamics, e.g., "saddle-to-saddle" (see, e.g., Jacot et al., 2021)
  - Simplicity bias (Shah et al., 2020; Boursier and Flammarion, 2024)
- **Need for simplified (yet relevant) models and architectures**
  - Quantitative (asymptotic or non-asymptotic) analysis
  - Empirical validity beyond the model (synthetic or real data)
  - From "understanding" to proposing improvements

Towards quantitative convergence

- **Idea 1: Adding noise to the dynamics**

$$W_k = W_{k-1} - \gamma \nabla F(W_{k-1}) + \sqrt{2\gamma\tau} \cdot \mathcal{N}(0, I)$$

  - Mei et al. (2018); Chizat (2022); Nitanda et al. (2022)
  - Allows for global quantitative convergence guarantees
  - Slow convergence as a function of temperature τ

Towards quantitative convergence

- **Idea 2: Change scaling**
  - Du et al. (2018, 2019); Allen-Zhu et al. (2019); etc.
  - Equivalent to replacing $h = \frac{1}{m}\sum_{j=1}^{m}\Psi(w_j)$ by $h = \frac{\alpha}{m}\sum_{j=1}^{m}\Psi(w_j)$ for $\alpha \to +\infty$
  - Allows for global linear convergence guarantees for deep architectures

- **But...**
  - “Lazy” regime where neurons do not move (Chizat, Oyallon, and Bach, 2019)
  - linear method equivalent to neural tangent kernel (Jacot et al., 2018)
  - No feature learning, little real effect in deep learning (Bietti and Bach, 2021)

Towards quantitative convergence

- **Idea 3: Simplify architectures**
  - Linear neural networks (e.g., Gidel et al., 2019; Marion and Chizat, 2024)
  - Diagonal linear networks (Woodworth et al., 2020; Pesme et al., 2021)
  - Precise and insightful guarantees, hard to extend to non-linear architectures

- **Idea 4: Simplify data models**
  - Gaussian or uniform data in high dimension (Zdeborová and Krzakala, 2016; Ghorbani et al., 2021, etc.)
  - Multiple index models (Bietti, Bruna, and Pillaud-Vivien, 2023)
  - Orthogonal inputs (Boursier, Pillaud-Vivien, and Flammarion, 2022)
  - Weakly correlated inputs (Dana, Bach, and Pillaud-Vivien, 2025)

Weakly correlated inputs (Dana, Bach, and Pillaud-Vivien, 2025)

- **One-hidden layer with square loss**
  - Fixed number m of hidden neurons and number n of observations
  - Generic initialization

- **Simplifying assumptions**
  - Empirical covariance matrix close to diagonal
  - Random data with diagonal population covariance matrix and $d \gtrsim n^2$
  - No assumptions on labels (no data model)

- **Main result**: Global exponential convergence to interpolating network as soon as $m \gtrsim \log(n)$ with characteristic time n
  - Explicit behavior through local Polyak-Lojasiewicz argument
  - Open problem: results for $d \gtrsim n$ with additional assumptions

Optimization for ML: current research and open problems

- **Optimal scaling of parameters initializations and normalizations**
  - Extension to deep networks (Yang and Hu, 2021; Chizat and Netrapalli, 2024)
  - Analysis on simple non-linear models (Bietti et al., 2023; Glasgow et al., 2025)
- **Analysis of modern architectures** (Resnets, transformers, Mamba, etc.)
  - Proof of convergence and proposition of improvements
- **Getting quantitative with "scaling laws"**
  - How much data and compute and data are needed to achieve a given performance? (Kaplan et al., 2020; Hoffmann et al., 2022; Paquette et al., 2024)
  - Taking into account data heterogeneity (Kunstner and Bach, 2025)
- **Open problem:** Why two reasonably wide hidden layers suffice to robustly reach global optimum?

References

Ahmet Alacaoglu and Yura Malitsky. Stochastic variance reduction for variational inequality methods. In *Conference on Learning Theory*, 2022.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in Neural Information Processing Systems*, 32, 2019.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(1): 629–681, 2017.

Francis Bach, Julien Mairal, and Jean Ponce. Convex sparse matrix factorizations. Technical Report 0812.1869, arXiv, 2008.

P. Balamurugan and F. Bach. Stochastic variance reduction methods for saddle-point problems. Technical Report 01319293, HAL, 2016.

A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Y. Bengio, N. Le Roux, P. Vincent, O. Delalleau, and P. Marcotte. Convex neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

Alberto Bietti and Francis Bach. Deep equals shallow for ReLU networks in kernel regimes. In *International Conference on Learning Representations*, 2021.

Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning Gaussian multi-index models with gradient flow. *arXiv preprint arXiv:2310.19793*, 2023.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2): 223–311, 2018.

Etienne Boursier and Nicolas Flammarion. Simplicity bias and optimization threshold in two-layer ReLU networks. *arXiv preprint arXiv:2410.02348*, 2024.

Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. *Advances in Neural Information Processing Systems*, 35, 2022.

M. A. Cauchy. Méthode générale pour la résolution des systèmes d'équations simultanées. *Comptes rendus des séances de l'Académie des sciences*, 25(1):536–538, 1847.

Lénaïc Chizat. Mean-field Langevin dynamics: Exponential convergence and annealing. *Transactions on Machine Learning Research*, 2022.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, pages 3036–3046, 2018.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pages 1305–1338, 2020.

Lénaïc Chizat and Praneeth Netrapalli. The feature speed formula: a flexible approach to scale hyper-parameters of deep neural networks. *Advances in Neural Information Processing Systems*, 2024.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information* *processing systems*, 32, 2019.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. *Advances in Neural Information Processing Systems*, 2019.

Léo Dana, Francis Bach, and Loucas Pillaud-Vivien. Convergence of shallow ReLU networks on weakly interacting data. *arXiv preprint arXiv:2502.16977*, 2025.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, 2014.

Alexandre Défossez, Leon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *Transactions on Machine Learning Research*, 2020.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, 2019.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems*, 2018.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2), 2021.

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. *Advances in Neural Information Processing Systems*, 2019.

Margalit Glasgow, Denny Wu, and Joan Bruna. Mean-field neural network beyond finite time horizon. 2025.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6151–6159, 2017.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841, 2018.

Benjamin D. Haeffele and René Vidal. Global optimality in neural network training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7331–7339, 2017.

Hadrien Hendrikx, Francis Bach, and Laurent Massoulié. Asynchronous accelerated proximal stochastic gradient for strongly convex distributed finite sums. Technical Report 1901.09865, arXiv, 2019.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. An empirical analysis of compute-optimal large language model training. *Advances in Neural Information Processing Systems*, 2022.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8580–8589, 2018.

Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*, 2021.

Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, 2013.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Frederik Kunstner and Francis Bach. Scaling laws for gradient descent and sign descent for linear bigram models under zipf's law. *arXiv preprint arXiv:2505.19227*, 2025.

V. Kurkova and M. Sanguineti. Bounds on rates of variable-basis and neural-network approximation. *IEEE Transactions on Information Theory*, 47(6):2659–2665, Sep 2001.

G. Lan. An optimal randomized incremental gradient method. Technical Report 1507.02000, arXiv, 2015.

N. Le Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in Neural Information Processing Systems*, 31, 2018.

H. Lin, J. Mairal, and Z. Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

Pierre Marion and Lénaïc Chizat. Deep linear networks for regression are implicitly regularized towards flat minima. In *Advances in Neural Information Processing Systems*, 2024.

Song Mei, Andrea Montanari, and P. Nguyen. A mean field view of the landscape of two-layers neural networks. *Proceedings of the National Academy of Sciences*, 115:E7665–E7671, 2018.

Y. Nesterov. A method for solving a convex programming problem with rate of convergence $O(1/k^2)$. *Soviet Math. Doklady*, 269

(3):543–547, 1983.

Y. Nesterov. *Introductory lectures on convex optimization: a basic course*. Kluwer, 2004.

Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, 2017.

A. Nitanda. Stochastic proximal gradient descent with acceleration techniques. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Atsushi Nitanda and Taiji Suzuki. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Convex analysis of the mean field Langevin dynamics. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+3 phases of compute-optimal neural scaling laws. In *Advances in Neural Information Processing Systems*, 2024.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 2021.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407, 1951.

S. Rosset, G. Swirszcz, N. Srebro, and J. Zhu. ℓ_1-regularization in infinite dimensional feature spaces. In *Proceedings of the Conference on Learning Theory (COLT)*, 2007.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 2020.

S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In *Proc. ICML*, 2014.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks. *arXiv preprint arXiv:1805.01053*, 2018.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, 2020.

Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, 2021.

Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5): 453–552, 2016.

L. Zhang, M. Mahdavi, and R. Jin. Linear convergence with condition number independent access of full gradients. In *Advances in Neural Information Processing Systems*, 2013.