Solving Overparametrized Systems of Random Equations

Andrea Montanari

Stanford University

June 26, 2024

The optimization puzzle in modern machine learning

- Empirical Risk Minimization (ERM) is highly non-convex
- Gradient methods find global optima

Working hypothesis

ERM becomes 'easy' if sufficiently overparametrized

The optimization puzzle in modern machine learning

- Empirical Risk Minimization (ERM) is highly non-convex
- Gradient methods find global optima

Working hypothesis

ERM becomes 'easy' if sufficiently overparametrized

The optimization puzzle in modern machine learning

Working hypothesis

ERM becomes 'easy' if sufficiently overparametrized

Can we understand this in a simple model?

Outline

1. A simple model with a long history
2. Gradient descent: Local analysis
3. Hessian descent
4. Exact solutions

Eliran Subag

Weizmann Institute

A small experiment with a small neural net

An experiment: 2-Layer ELU network

$$f(\boldsymbol{z}; \boldsymbol{W}) = \frac{a}{\sqrt{m}} \sum_{j=1}^{m} s_i \, \sigma(\langle \boldsymbol{w}_j, \boldsymbol{z} \rangle) \ , \qquad \boldsymbol{z} \in \mathbb{R}^D \, .$$

$$\sigma(x) = \begin{cases} x & \text{if } x \geq 0, \\ e^x - 1 & \text{if } x < 0. \end{cases} , \qquad \|\boldsymbol{W}\|_F^2 = \sum_{i=1}^{m} \|\boldsymbol{w}_i\|_2^2 \leq m \, .$$

Empirical Risk Minimization via SGD

$$R_n(\boldsymbol{W}) := \frac{1}{2n} \sum_{i=1}^{n} \big(y_i - f(\boldsymbol{z}_i; \boldsymbol{W})\big)^2 ,$$

$$\widetilde{\boldsymbol{W}}^{k+1} = \boldsymbol{W}^k - \frac{\eta_k}{2} \sum_{i \in B(k)} \nabla_{\boldsymbol{w}} \big(y_i - f(\boldsymbol{z}_i; \boldsymbol{W}^k)\big)^2 ,$$

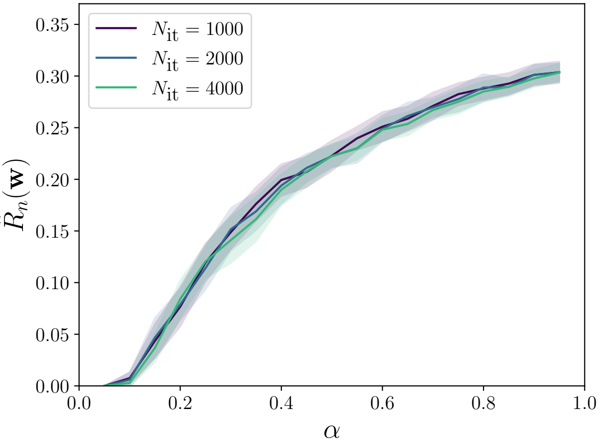
$$\boldsymbol{W}^{k+1} = \mathsf{Proj}(\widetilde{\boldsymbol{W}}^{k+1}) .$$

$\eta_k = \mathrm{lr}/(1 + \mathrm{epoch}(k))^{1/2}$ $\boldsymbol{W}^0 \sim \mathsf{N}(0, \varepsilon^2 \mathbf{I}_{mD}/D)$, $\varepsilon = 0.03$

Data distribution

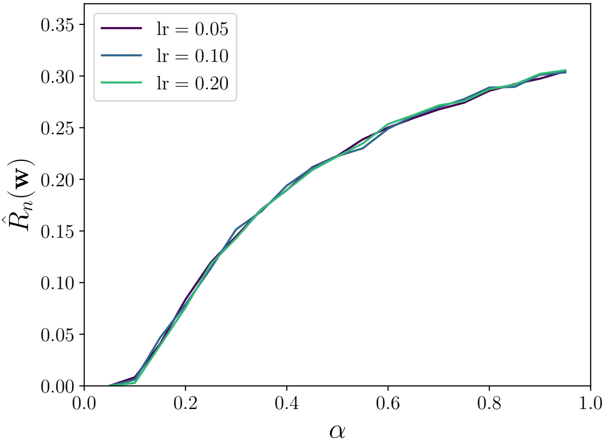
$$(\boldsymbol{z}_i, y_i) \sim \mathsf{N}(0, \mathbf{I}_D) \otimes \mathrm{Unif}(\{+1, -1\})$$

Varying number of epochs; $\alpha = n/mD$



- $m = D = 20$, lr $= 0.1$
- Averages over 20 realizations; one std bands

Varying learning rate; $\alpha = n/mD$



- $m = D = 20$, $N_{\rm it} = 2,000$
- Averages over 20 realizations

A simple model with a long history

General problem

1. $F_1, \ldots, F_n : \mathbb{R}^d \to \mathbb{R}$ i.i.d. random functions.

2. Find $\boldsymbol{x} \in \mathbb{S}^{d-1}$ such that $F_i(\boldsymbol{x}) = 0$ for all $i \leq n$.

Gaussian model

F_i centered, rotation invariant (in distr) Gaussian processes

- Covariance
$$\mathbb{E}\left[F_i(\boldsymbol{x}^1)F_j(\boldsymbol{x}^2)\right] = \delta_{ij}\xi\left(\langle \boldsymbol{x}^1, \boldsymbol{x}^2\rangle\right).$$

- $\mathbf{F}(\boldsymbol{x}) = (F_1(\boldsymbol{x}), \ldots, F_n(\boldsymbol{x}))^{\mathsf{T}}$.

Cost function and approximate solutions

$$R_n(x) := \frac{1}{2}\|F(x)\|_2^2$$

- At a fixed $x_0 \in \mathbb{S}^{d-1}$, $R_n(x_0) = n\xi(1)/2 + o(n)$
- Solutions

$$\mathsf{Sol}_{n,d}(\varepsilon) := \left\{x \in \mathbb{S}^{d-1} : \ \|F(x)\|_2^2 \leq n\xi(1) \cdot \varepsilon\right\}.$$

- $n, d \to \infty$, $n/d \to \alpha$.

Questions

Q1 Do exact solutions exist: $\mathsf{Sol}_{n,d}(0) \neq \emptyset$?

Do approximate solution exist, $\mathsf{Sol}_{n,d}(\varepsilon) \neq \emptyset$?

Q2 Can we find them in polytime?

Questions

Q1 Do exact solutions exist: $\mathsf{Sol}_{n,d}(0) \neq \emptyset$?

Do approximate solution exist, $\mathsf{Sol}_{n,d}(\varepsilon) \neq \emptyset$?

Q2 **Can we find them in polytime?**

History: Classical (complex) setting

$$F : \mathbb{C}^d \to \mathbb{C}^n, \quad \text{homogeneous, } \deg(F_i) = p_i$$

Q1 Bezout's theorem (1779)

For $n = d - 1$, deterministically:

$$|\mathsf{Sol}_{n,d}(0)| = \prod_{i=1}^{n} p_i$$

Q2

- **Smale 17th problem** (1993-1998)
- Positive answer (Lairez, 2020)
- Homotopy methods

History: Classical (complex) setting

$$F : \mathbb{C}^d \to \mathbb{C}^n, \quad \text{homogeneous, } \deg(F_i) = p_i$$

Q1 Bezout's theorem (1779)

For $n = d - 1$, deterministically:

$$|\mathrm{Sol}_{n,d}(0)| = \prod_{i=1}^{n} p_i$$

Q2

- **Smale 17th problem** (1993-1998)
- Positive answer (Lairez, 2020)
- Homotopy methods

History: Classical (complex) setting

$$\mathsf{F} : \mathbb{C}^d \to \mathbb{C}^n, \quad \text{homogeneous, } \deg(\mathsf{F}_i) = p_i$$

Q1 Bezout's theorem (1779)

For $n = d - 1$, deterministically:

$$|\mathsf{Sol}_{n,d}(0)| = \prod_{i=1}^{n} p_i$$

Q2

- **Smale 17th problem** (1993-1998)
- Positive answer (Lairez, 2020)
- Homotopy methods

History: Real setting

Q1 Homogeneous case, $n = d - 1$.

- Subag 2022: With high probability

$$|\mathsf{Sol}_{n,d}(0)| = (1 + o(1)) \prod_{i=1}^{n} \sqrt{p_i}$$

Q1 Non-homogeneous case, $n/d \to \alpha \in (0, \infty)$.

- Next slide

Q2 **The rest of this talk**

History: Real setting

Q1 Homogeneous case, $n = d - 1$.

- Subag 2022: With high probability

$$|\mathsf{Sol}_{n,d}(0)| = (1 + o(1)) \prod_{i=1}^{n} \sqrt{p_i}$$

Q1 Non-homogeneous case, $n/d \to \alpha \in (0, \infty)$.

- Next slide

Q2 **The rest of this talk**

History: Real setting

Q1 Homogeneous case, $n = d - 1$.

- Subag 2022: With high probability

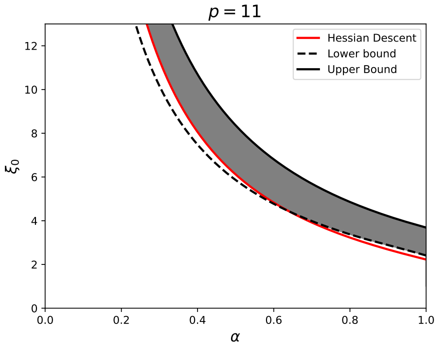
$$|\mathsf{Sol}_{n,d}(0)| = (1 + o(1)) \prod_{i=1}^{n} \sqrt{p_i}$$

Q1 Non-homogeneous case, $n/d \to \alpha \in (0, \infty)$.

- Next slide

Q2 **The rest of this talk**

Q1: $\xi(q) = \xi_0 + q^{11}$ (polynomial of degree 11)



- Above gray region, $\alpha > \alpha_{\mathrm{UB}}(\xi_0)$: $\mathsf{Sol}_{n,d}(\varepsilon) = \emptyset$
- Below gray region, $\alpha < \alpha_{\mathrm{LB}}(\xi_0)$: $\mathsf{Sol}_{n,d}(0) = \emptyset$

See paper for formal statements.

Gradient descent: Local analysis

Projected Gradient Descent

Gradient descent

$$
\begin{aligned}
\boldsymbol{z}^{k+1} &= \boldsymbol{x}^{k} - \eta \mathsf{P}_{\mathsf{T},\boldsymbol{x}^{k}} \nabla \mathsf{R}_{n}(\boldsymbol{x}^{k}) \,, \\
\boldsymbol{x}^{k+1} &= \frac{\boldsymbol{z}^{k+1}}{\|\boldsymbol{z}^{k+1}\|_{2}} \,.
\end{aligned}
$$

Projected gradient flow

$$
\dot{\boldsymbol{x}}(t) = -\mathsf{P}_{\mathsf{T},\boldsymbol{x}(t)} \nabla \mathsf{R}_{n}(\boldsymbol{x}(t)) \,.
$$

Difficult to analyze sharply!

Projected Gradient Descent

Gradient descent

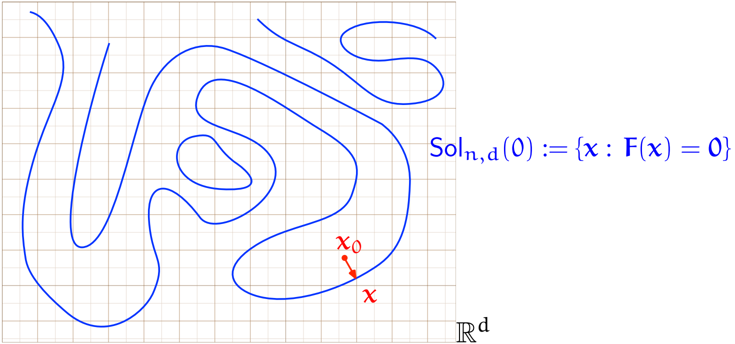
$$\begin{aligned} \boldsymbol{z}^{k+1} &= \boldsymbol{x}^{k} - \eta \mathsf{P}_{\mathsf{T},\boldsymbol{x}^{k}} \nabla \mathsf{R}_{n}(\boldsymbol{x}^{k}) \,, \\ \boldsymbol{x}^{k+1} &= \frac{\boldsymbol{z}^{k+1}}{\|\boldsymbol{z}^{k+1}\|_2} \,. \end{aligned}$$

Projected gradient flow

$$\dot{\boldsymbol{x}}(t) = -\mathsf{P}_{\mathsf{T},\boldsymbol{x}(t)} \nabla \mathsf{R}_{n}(\boldsymbol{x}(t)) \,.$$

Difficult to analyze sharply!

Local analysis: Taylor expand around initialization



State of the art in ML Theory: Jacot, Gabriel, Hongler, 2018; Du, Zhai, Poczos, Singh 2018; Allen-Zhu, Li, Song 2018; Chizat, Bach, 2019; Arora, Du, Hu, Li, Salakhutdinov, Wang, 2019; Oymak, Soltanolkotabi, 2019; . . .

Local analysis

$$\underline{\alpha}_{\mathrm{GD}}(\xi) := \frac{c_0 \xi'(1)^2}{\xi''(1)\xi(1)\big(\log(\xi'''(1)/\xi''(1)) \vee 1\big)}\,.$$

Theorem (M, Subag, 2023)

If $\alpha < \underline{\alpha}_{\mathrm{GD}}(\xi)$*, and* $\eta < 1/(C_1 d)$*, then whp for all* $k \geq 1$,

$$\|F(x^k)\|_2^2 \leq 2n\xi(1)\, e^{-c_2(\sqrt{d}-\sqrt{n})^2(\eta k)}\,.$$

Special case: $\xi(q) = \xi_0 + q^p$

$$\underline{\alpha}_{\mathrm{GD}}(\xi_0, p) \asymp \frac{1}{\xi_0 \log p}\,.$$

To be compared with

$$\alpha_{\mathrm{LB}}(\xi_0, p) = \frac{\log p}{\xi_0} \cdot \big(1 + o_p(1)\big)\,.$$

Hessian descent

Problem with GD

- Need $\|\nabla R_n(\boldsymbol{x}^k)\|_2 \geq \varepsilon$.
- Cannot be true uniformly.
- $\Rightarrow$ Local analysis :(

Problem with GD

- Need $\|\nabla R_n(\boldsymbol{x}^k)\|_2 \geq \varepsilon$.
- Cannot be true uniformly.
- $\Rightarrow$ Local analysis :(

Problem with GD

- Need $\|\nabla R_n(\boldsymbol{x}^k)\|_2 \geq \varepsilon$.
- Cannot be true uniformly.
- $\Rightarrow$ Local analysis :(

Problem with GD

- Need $\|\nabla R_n(\boldsymbol{x}^k)\|_2 \geq \varepsilon$.
- Cannot be true uniformly.
- $\Rightarrow$ Local analysis :(

Idea: Use Hessian

Problem with Hessian Descent

$$\mathbf{D}^2 \mathsf{R}_n(\boldsymbol{x}) = \nabla^2 \mathsf{R}_n(\boldsymbol{x})|_{\mathsf{T}(\boldsymbol{x})} - \langle \boldsymbol{x}, \nabla \mathsf{R}_n(\boldsymbol{x}) \rangle \mathbf{I}$$

- $\mathbf{D}^2$ = Riemannian Hessian
- ∇^2 = Euclidean Hessian
- $\mathsf{T}(\boldsymbol{x})$ = Tangent space at $\boldsymbol{x}$
- Problem: $\langle \boldsymbol{x}, \nabla \mathrm{R}_n(\boldsymbol{x}) \rangle$ does not concentrate

Problem with Hessian Descent

$$\mathbf{D}^2 \mathsf{R}_n(\boldsymbol{x}) = \nabla^2 \mathsf{R}_n(\boldsymbol{x})|_{\mathsf{T}(\boldsymbol{x})} - \langle \boldsymbol{x}, \nabla \mathsf{R}_n(\boldsymbol{x}) \rangle \mathbf{I}$$

- $\mathbf{D}^2$ = Riemannian Hessian
- ∇^2 = Euclidean Hessian
- $\mathsf{T}(\boldsymbol{x})$ = Tangent space at $\boldsymbol{x}$
- Problem: $\langle \boldsymbol{x}, \nabla \mathsf{R}_n(\boldsymbol{x}) \rangle$ does not concentrate

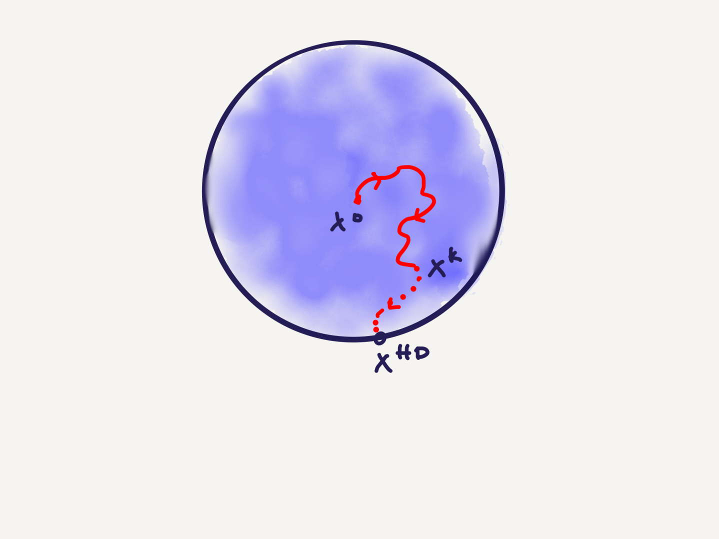
Problem with Hessian Descent

$$\mathbf{D}^2 R_n(x) = \nabla^2 R_n(x)|_{T(x)} - \cancel{\langle x, \nabla R_n(x)\rangle \mathbf{I}}$$

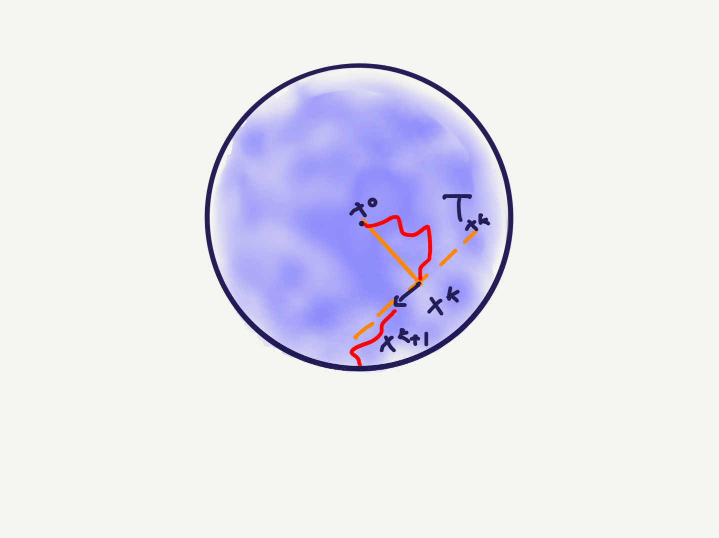
Idea: Relax sphere constraint

Subag, 2020 (spherical spin glasses)

Algorithm: Sketch



Algorithm: Orthogonal steps



Algorithm: Simplified

$$\boldsymbol{v}_{\min}(\boldsymbol{A}) := \text{eigenvector associated to smallest eigenvalue of } \boldsymbol{A}$$

Initialize $\boldsymbol{x}^1 \sim \sqrt{\delta} \cdot \mathrm{Unif}(\mathbb{S}^{d-1})$;
for $k \in \{1, \ldots, K := 1/\delta - 1\}$ **do**
  Compute $\boldsymbol{v}(\boldsymbol{x}^k) = \boldsymbol{v}_{\min}(\nabla^2 R_n(\boldsymbol{x}^k)|_{T_{\boldsymbol{x}^k}})$;
  $s_k := \mathrm{sign}(\langle \boldsymbol{v}(\boldsymbol{x}^k), \nabla R_n(\boldsymbol{x}^k) \rangle)$;
  $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - s_k \sqrt{\delta}\, \boldsymbol{v}(\boldsymbol{x}^k)$;
end
return $\boldsymbol{x}^{\mathrm{HD}} = \boldsymbol{x}^K$;

Full algorithm

Initialize $\boldsymbol{x}^1 \sim \sqrt{\delta} \cdot \mathrm{Unif}(\mathbb{S}^{d-1})$;
for $k \in \{1, \dots, K := 1/\delta - 1\}$ **do**

Compute $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{x}^k) \in \mathsf{T}_{\boldsymbol{x}^k}$ such that $\|\boldsymbol{v}\|_2 = 1$ and

$$\langle \boldsymbol{v}, \nabla^2 R_n(\boldsymbol{x}^k)\boldsymbol{v}\rangle \le \lambda_{\min}(\nabla^2 R_n(\boldsymbol{x}^k)|_{\mathsf{T},\boldsymbol{x}^k}) + d\delta\,;$$

$s_k := \mathrm{sign}(\langle \boldsymbol{v}(\boldsymbol{x}^k), \nabla R_n(\boldsymbol{x}^k)\rangle)$;
$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - s_k\sqrt{\delta}\,\boldsymbol{v}(\boldsymbol{x}^k)$;

end
return $\boldsymbol{x}^{\mathrm{HD}} = \boldsymbol{x}^K$;

Analysis

Theorem (M, Subag, 2023)

For $\alpha \in (0,1)$, $a, b \in \mathbb{R}_{\geq 0}$, *define*

$$z_*(\alpha, a, b) := \inf_{m>0} \left\{ \frac{1}{m} - \frac{\alpha b}{1+bm} + a^2 m \right\}.$$

Let $u(\,\cdot\,; \alpha, \xi) : [0,1] \to \mathbb{R}$ *be the unique solution of the ODE*

$$\frac{du}{dt}(t) = -\frac{1}{2\alpha} z_*\left(\alpha; \sqrt{2\alpha u(t)\xi''(t)}, \xi'(t)\right), \qquad u(0) = \frac{1}{2}\xi(0).$$

Then whp

$$\frac{1}{n} R_n(\boldsymbol{x}^{\mathrm{HD}}) \leq u(1; \alpha, \xi) + C_0 \delta.$$

Recall $R_n(\boldsymbol{x}) := \|\boldsymbol{F}(\boldsymbol{x}^{\mathrm{HD}})\|_2^2/2$.

Where does this come from?

Hessian

$$\nabla^2 R_n(x) = \sum_{\ell=1}^{n} F_\ell(x)\nabla^2 F_\ell(x) + \mathbf{D}\mathbf{F}(x)^\mathsf{T}\mathbf{D}\mathbf{F}(x)$$

Distribution as x, $\|x\|^2 = q$

$$\nabla^2 R_n(x)|_{T(x)} = \sqrt{R_n(x)\xi''(q)}\,\boldsymbol{W} + \xi'(q)\,\mathbf{Z}^\mathsf{T}\mathbf{Z}\,,$$

$$(\boldsymbol{W}, \mathbf{Z}) \sim \mathrm{GOE}(d-1) \otimes \mathrm{GOE}(n, d-1)$$

Decrease in value

$$\lim_{n,d\to\infty} \frac{1}{d}\lambda_{\min}(\boldsymbol{A}_{n,d}) = -z_*(\alpha, a, b) := -\inf_{m>0} \frac{1}{m} - \frac{\alpha b}{1+bm} + a^2 m\,.$$

Where does this come from?

Hessian

$$\nabla^2 R_n(\boldsymbol{x}) = \sum_{\ell=1}^{n} F_\ell(\boldsymbol{x}) \nabla^2 F_\ell(\boldsymbol{x}) + \mathbf{DF}(\boldsymbol{x})^{\mathsf{T}} \mathbf{DF}(\boldsymbol{x})$$

Distribution as x, $\|x\|^2 = q$

$$\nabla^2 R_n(x)|_{\mathsf{T}(x)} = \sqrt{R_n(x)\xi''(q)}\, \boldsymbol{W} + \xi'(q)\, \mathbf{Z}^{\mathsf{T}}\mathbf{Z}\,,$$

$$(\boldsymbol{W}, \mathbf{Z}) \sim \mathrm{GOE}(d-1) \otimes \mathrm{GOE}(n, d-1)$$

Decrease in value

$$\lim_{n,d\to\infty} \frac{1}{d} \lambda_{\min}(\boldsymbol{A}_{n,d}) = -z_*(\alpha, a, b) := -\inf_{m>0} \frac{1}{m} - \frac{\alpha b}{1+bm} + a^2 m\,.$$

Where does this come from?

Hessian

$$\nabla^2 R_n(x) = \sum_{\ell=1}^{n} F_\ell(x)\nabla^2 F_\ell(x) + \mathbf{D}\mathbf{F}(x)^\mathsf{T}\mathbf{D}\mathbf{F}(x)$$

Distribution as x, $\|x\|^2 = q$

$$\nabla^2 R_n(x)|_{\mathsf{T}(x)} = \sqrt{R_n(x)\xi''(q)}\,\boldsymbol{W} + \xi'(q)\,\mathbf{Z}^\mathsf{T}\mathbf{Z}\,,$$

$$(\boldsymbol{W}, \mathbf{Z}) \sim \mathrm{GOE}(d-1) \otimes \mathrm{GOE}(n, d-1)$$

Decrease in value

$$\lim_{n,d\to\infty} \frac{1}{d}\lambda_{\min}(A_{n,d}) = -z_*(\alpha, a, b) := -\inf_{m>0} \frac{1}{m} - \frac{\alpha b}{1+bm} + a^2 m\,.$$

Where does this come from?

Hessian

$$\nabla^2 R_n(\boldsymbol{x}) = \sum_{\ell=1}^{n} F_\ell(\boldsymbol{x}) \nabla^2 F_\ell(\boldsymbol{x}) + \mathbf{DF}(\boldsymbol{x})^\mathsf{T} \mathbf{DF}(\boldsymbol{x})$$

Distribution as $\boldsymbol{x}$, $\|\boldsymbol{x}\|^2 = q$

$$\nabla^2 R_n(\boldsymbol{x})|_{\mathsf{T}(\boldsymbol{x})} = \sqrt{R_n(\boldsymbol{x})\xi''(q)}\, \boldsymbol{W} + \xi'(q)\, \mathbf{Z}^\mathsf{T}\mathbf{Z}\,,$$

$$(\boldsymbol{W}, \mathbf{Z}) \sim \mathrm{GOE}(d-1) \otimes \mathrm{GOE}(n, d-1)$$

Decrease in value

$$\lim_{n,d\to\infty} \frac{1}{d} \lambda_{\min}(\boldsymbol{A}_{n,d}) = -z_*(\alpha, a, b) := -\inf_{m>0} \frac{1}{m} - \frac{\alpha b}{1+bm} + a^2 m\,.$$

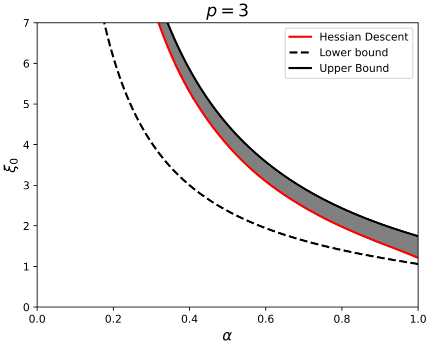
Special case: $\xi(q) = \xi_0 + q^p$

$$\frac{4(p-1)}{p\xi_0 + 4(p-1)} \le \alpha_{\rm HD}(\xi_0, p) \le \frac{4(p-1)}{p\xi_0}\,.$$

To be compared with

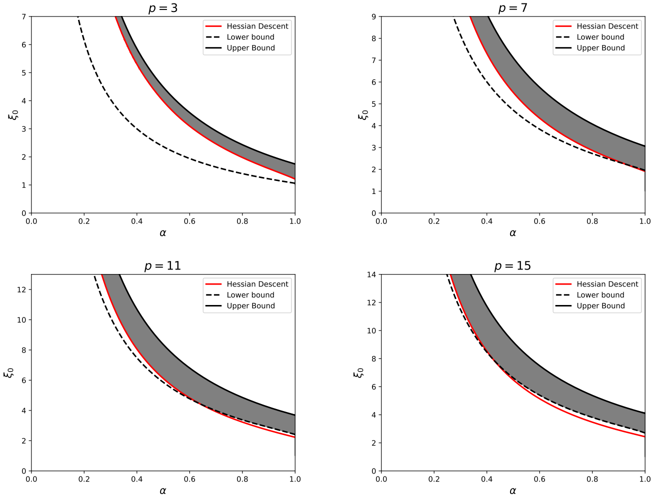
$$\underline{\alpha}_{\rm GD}(\xi_0, p) \asymp \frac{1}{\xi_0 \log p}\,,$$
$$\alpha_{\rm LB}(\xi_0, p) = \frac{\log p}{\xi_0} \cdot \big(1 + o_p(1)\big)\,.$$

Phase diagram ($\xi(q) = \xi_0 + q^3$)



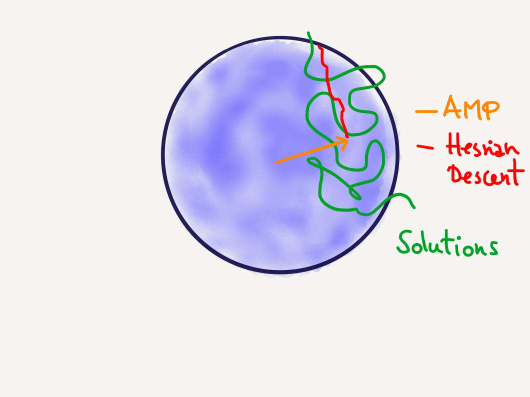
- Above gray region, $\alpha > \alpha_{\mathrm{UB}}(\xi_0)$: $\mathsf{Sol}_{n,d}(\varepsilon) = \emptyset$
- Below gray region, $\alpha < \alpha_{\mathrm{LB}}(\xi_0)$: $\mathsf{Sol}_{n,d}(0) = \emptyset$
- Red line: $\alpha_{\mathrm{HD}}(\xi_0, p)$

Phase diagram ($\xi(\mathbf{q}) = \xi_0 + \mathbf{q}^p$)



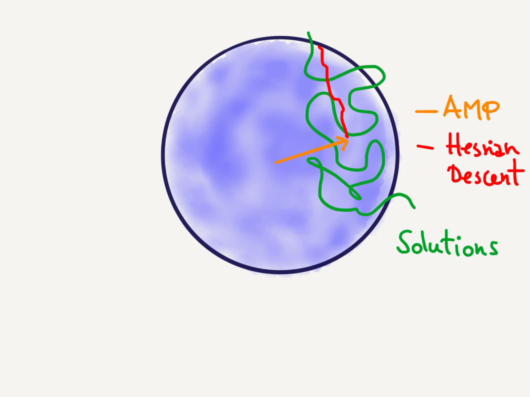
Is HD optimal (among polytime algs)?

- **No!** Suboptimal when $\mathsf{F}(\boldsymbol{x})$ has degree-1 term
- $\mathsf{Sol}_{n,d}(0)$ not centered at $\mathbf{0}$
- See paper for the fix/general algorithm
- Conjectured to be optimal among 'stable algorithms'



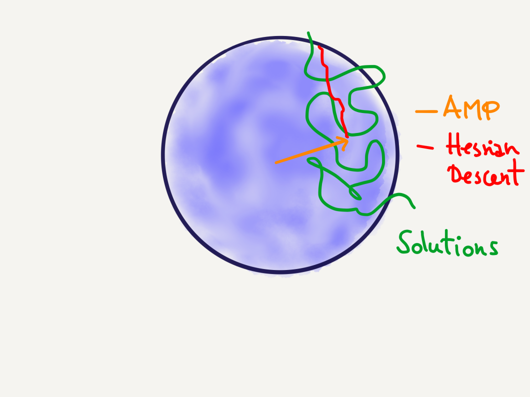
Is HD optimal (among polytime algs)?

- **No!** Suboptimal when $\mathsf{F}(\boldsymbol{x})$ has degree-1 term
- $\mathsf{Sol}_{n,d}(0)$ not centered at $\mathbf{0}$
- See paper for the fix/general algorithm
- Conjectured to be optimal among 'stable algorithms'



Is HD optimal (among polytime algs)?

- **No!** Suboptimal when $\mathsf{F}(\boldsymbol{x})$ has degree-1 term
- $\mathsf{Sol}_{n,d}(0)$ not centered at $\mathbf{0}$
- See paper for the fix/general algorithm
- Conjectured to be optimal among 'stable algorithms'



What about exact solutions?

What is an exact solution?

Definition (Shub, Smale, 1993)

$\mathbf{x}_*$ is an *approximate solution* of $F(\mathbf{x}) = 0$ if letting $(\mathbf{x}^k)_{k\geq 0}$ be Newton iterates with $\mathbf{x}^0 = \mathbf{x}_*$, then, for all k

$$\|F(\mathbf{x}^k)\| \leq \|F(\mathbf{x}^0)\| \cdot \exp\left\{-c \cdot 2^k\right\}.$$

Smale 17th problem over the reals:

Can we find approximate solutions in polytime?

What is an exact solution?

Definition (Shub, Smale, 1993)

$\mathbf{x}_*$ is an *approximate solution* of $F(\mathbf{x}) = 0$ if letting $(\mathbf{x}^k)_{k \geq 0}$ be Newton iterates with $\mathbf{x}^0 = \mathbf{x}_*$, then, for all k

$$\|F(\mathbf{x}^k)\| \leq \|F(\mathbf{x}^0)\| \cdot \exp\left\{-c \cdot 2^k\right\}.$$

Smale 17th problem over the reals:

Can we find approximate solutions in polytime?

What is an exact solution?

Theorem (M, Subag, 2024)

Assume F_i homogeneous, arbitrary (possibly different) degrees. Then there exists a deterministic polytime algorithm such that, if

$$n \leq d - C\sqrt{d \log d},$$

then it return a an approximate solution, with high probability wrt $\mathbf{F}$.

Conclusion #1

- Random systems of nonlinear equations
- Rich computational/probabilistic structure
- Quantitative comparison with neural nets lanscape?

Conclusion #1

- Random systems of nonlinear equations
- Rich computational/probabilistic structure
- Quantitative comparison with neural nets lanscape?

Conclusion #2

It is an honor to celebrate Andrew!
Thanks!

Epilogue: Revisiting the original experiment

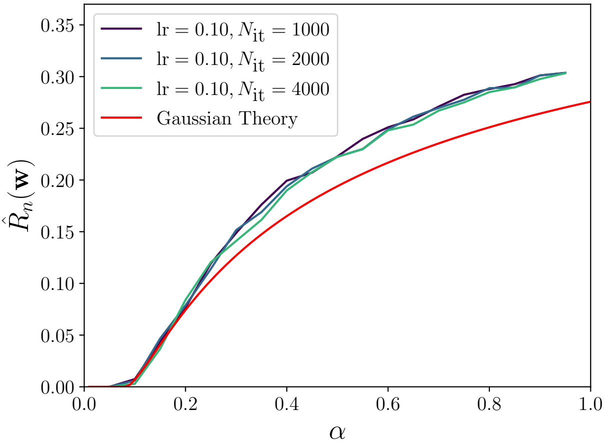
Empirical Risk Minimization

$$f(\boldsymbol{z};\boldsymbol{W}) = \frac{a}{\sqrt{m}} \sum_{j=1}^{m} s_i\, \sigma(\langle \boldsymbol{w}_j, \boldsymbol{z} \rangle) \,, \qquad \boldsymbol{z} \in \mathbb{R}^D \,.$$

$$R_n(\boldsymbol{W}) := \frac{1}{2n} \sum_{i=1}^{n} \big(y_i - f(\boldsymbol{z}_i;\boldsymbol{W})\big)^2 \,,$$

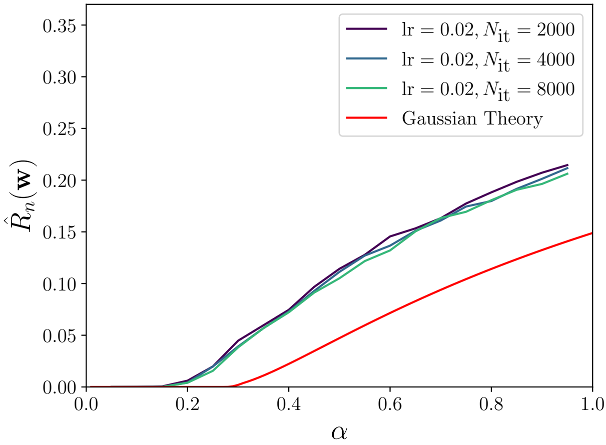
$$\|\boldsymbol{W}\|_F^2 \leq m \,.$$

Experiments vs Gaussian Theory: $\alpha = 1$



Red: Approx matching covariance

Experiments vs Gaussian Theory: $\alpha = 2$



Experiments vs Gaussian Theory: $\alpha = 5$