Algorithmic Regularization in Over-parameterized Models

Tengyu Ma
Facebook AI Research

Based on joint work with Yuanzhi Li (Princeton) and Hongyang Zhang (Stanford)
Over-parameterization: \# parameters \(\gg\) \# examples

\(\exists\) a set of parameters that can

- fit to training data and generalize to test data
- or fit to real inputs with random labels, and fail to generalize
- or fit to training data but fail to generalize

This talk: analysis for simpler models that share the properties above (matrix sensing, and neural nets with quadratic activations)
Explaining Generalization?

- Uniform convergence doesn’t hold
  - training loss \( \approx \) test loss for all parameters
  - \( \exists \) a model that can fit to training data but fail to generalize

- Algorithm matters: multiple local/global minima exist, the algorithm chooses the one that generalizes
  - different algorithms converge to local min of training loss, but generalize differently [Keskar et al’16, Wilson et al’17, Dinh et al’17]

- Post-mortem explanations: margin theory, PAC-Bayes, and compression-based bounds [Bartlett et al.’17, Neyshabur et al.’17, Arora et al.’18, Dziugaite and Roy’18]
Stochastic gradient descent, with proper initialization and learning rate, prefers an optimal solution with low complexity, when it exists.

- Algorithms matter
- Intrinsic complexity of the data matters

- # parameters is almost irrelevant
- This talk: rigorous argument for matrix sensing and quadratic neural networks
Warm-up: Linear Models

- $n$ data $x_1, \ldots, x_n \in \mathbb{R}^d$, $d \gg n$

$$\min_{\beta} f(\beta) := \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2$$

- Claim: minimize $f(\beta)$ with gradient descent starting with $\beta = 0$ is equivalent to solve

$$\min_{\beta} \|\beta\|$$

subject to

$$\sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 = 0$$

- Gradient descent is limited to search in a subspace

- Related: GD on logistic loss converges to max margin solution [Soudry et al.’17, Ji&Telgarsky’17]
Matrix Sensing with Over-parameterization

- $n$ data points $A_1, ..., A_n \in \mathbb{R}^{d \times d}$ from standard normal dist.
- Unknown PSD matrix $M \in \mathbb{R}^{d \times d}$ of rank $r \ll d$
- We observe
  \[ y_i = \langle A_i, M \rangle \]
- Variable $U \in \mathbb{R}^{d \times d}$
  \[ \min_U f(U) = \sum_{i=1}^{n} (y_i - \langle A_i, UU^\top \rangle)^2 \]
- Focus: gradient descent $U_{t+1} = U_t - \eta \nabla f(U_t)$
- Well-studied problem with efficient solutions [Recht et al.’10, Candes et al’07, Tu et al’2015, Zheng and Lafferty’15...]
Matrix Sensing with Over-parameterization (Cont’d)

\[ y_i := \langle A_i, M \rangle \]

\[ \min_U f(U) = \sum_{i=1}^{n} (y_i - \langle A_i, UU^\top \rangle)^2 \]

- Regime of parameters: \( n \approx d r^2 \ll d^2 \)
- Ideal solution: \( U \) satisfying \( UU^\top = M \) has zero training error
- \( \exists \) other solution \( U \) with zero training error but \( UU^\top \neq M \)

Gradient descent with small initialization empirically converges to the ideal solution! [Gunasekar et al.’2017]

- Compared to low-rank factorization (taking \( U \in \mathbb{R}^{d \times r} \)): the algorithm finds the correct rank automatically
test error (population risk) = $\mathbb{E} f = \|M - UU^\top\|^2_F$

- $r = 5, n = 5d r$
- Early stopping and stochasticity is not necessary

Systematic empirical studies in [Gunasekar et al.’2017]
**Main Results**

**Theorem:** [Li-M.-Zhang’17] With $\tilde{O}(dr^2)$ observations, and initialization $U_0 = \alpha \cdot I$ and learning rate $\eta$, at iteration $T$ satisfying

$$\frac{1}{\eta} \cdot \log \frac{d}{\alpha} \lesssim T \lesssim \frac{1}{\eta} \cdot \frac{1}{\sqrt{d\alpha}},$$

the generalization error is bounded by

$$\|U_T U_T^\top - M\|_F^2 \lesssim d\alpha$$

**Technicalities:**

- We assume $M$ is well-conditioned
- Theorem also holds when the measurements $A_1, \ldots, A_n$ satisfy $\delta$-restricted isometry property with $\delta \lesssim 1/\sqrt{r}$
- The runtime bound is non-trivial even with infinite samples
Key Intuitions

Gradient descent prefers “low complexity” solutions

\[ S_r = \{ \text{approximately rank-r solutions} \} \]
\[ := \{ U : \sigma_{r+1}(U) \leq \epsilon \} \]

Non-generalizable global minima of training loss

generalizable global minima of training loss
More concrete analysis plan:

- GD on population risk $\mathbb{E}f$ stays in $S_r$
- GD on $f$ behaves similarly to that on $\mathbb{E}f$ in $S_r$ \( \nabla \mathbb{E}f(U) \approx \nabla f(U) \)
- Generalization is trivial in $S_r$

$$\mathbb{E}f(U) \approx f(U), \forall U \in S_r$$
Weight matrix has low-complexity throughout the training

- Input dim = 100
- Generate labels with a network of hidden layer size $r = 1$
- Train with hidden layer size = 100
Proof Sketch for Rank-1 Case:

- WLOG, assume $M = u^*u^{*\top}$, $\|u^*\| = 1$
- Decompose the iterate $U_t$ into:
  \[ U_t = u^*r_t^\top + E_t \]
  \[
  \text{signal} \quad \text{noise}
  \]

Goals: show inductively
- $\|E_t\|_{\text{op}} \approx 0$
- $\|r_t\| \to 1$
- These imply $U_t \to u^*u^{*\top}$

$\delta \approx \frac{dr^2}{n}$

\[ \|E_{t+1}\| \leq \|E_t\| + 2\eta\delta \]
\[ \|r_{t+1}\| \geq (1 + \eta(1 - \|r_t\|))\|r_t\| - 2\eta\delta \]
Lemma 1: $\|E_{t+1}\| \leq \|E_t\| + 2\eta \delta$

- **Preparation:**
  \[
  U_{t+1} = U_t - \eta \nabla f(U_t)
  \]
  \[
  = U_t - \eta (U_t U_t - M) U_t
  \]
  Small when $U_t$ is approx. low rank

- **Proof:**
  \[
  E_t = (I - u^* u^*) U_t
  \]
  \[
  E_{t+1} = E_t (I - \eta U_t U_t) + \text{small term}
  \]
  \[
  \text{b.c.} \quad (I - u^* u^*) M = 0
  \]
  GD on population risk reduces the error
  \[
  \|E_{t+1}\| \leq \|E_t\| + \text{small term}
  \]
Quadratic Neural Nets

- $q(\cdot)$: entry-wise quadratic
  $$y = 1^T \cdot q(U^Tx)$$

- Almost equivalent to matrix sensing with rank-1 measurement:
  $$y = \|U^Tx\|^2 = \langle xx^T, \ UU^T \rangle$$

- Only difference: unlike random measurement, $xx^T$ doesn’t satisfy restricted isometry property

- Solution: throwing away a very small fraction of the data (adaptively) that devastate restricted isometry property
Simulations (for Matrix Sensing)

- Generalization error depends on initialization

Initialization = $\alpha \cdot I$
Initialization = \( I \)

- Caveat: SGD or GD with large initialization **can** work with quadratic neural networks. (But the current theory requires small initialization.)
Simulations (Cont’d): Factorized Parameterization Matters

- Algo. analyzed: GD on
  \[ \min_U f(U) = \sum_{i=1}^{n} (y_i - \langle A_i, U U^\top \rangle)^2 \]

- Algo. for comparison: projected GD on
  \[ \min_{Z \succeq 0} g(Z) = \sum_{i=1}^{n} (y_i - \langle A_i, Z \rangle)^2 \]
Conclusion

- Algorithms have an implicit regularization effect

Open questions:

- other matrix factorization based models
  - logistic loss [Gunasekar et al’18]
- neural nets with other activation functions and loss (more in Nati’s talk)
- better understanding of algorithms for deep learning
  - which seems to be very helpful for fully understanding generalization

Thank you!