Overparametrization for Landscape Design in Non-convex Optimization

Jason D. Lee

University of Southern California

October 8, 2018

The State of Non-Convex Optimization

- Practical observation: Empirically, non-convexity is not an issue. Gradient methods find high quality solutions.

The State of Non-Convex Optimization

- Practical observation: Empirically, non-convexity is not an issue. Gradient methods find high quality solutions.
- Theoretical Side: NP-hard in many cases. *e.g.* Finding a local minimum in a smooth function is NP-Hard. Finding a 2nd order optimal point in a non-smooth function is NP-Hard.

The State of Non-Convex Optimization

- Practical observation: Empirically, non-convexity is not an issue. Gradient methods find high quality solutions.
- Theoretical Side: NP-hard in many cases. *e.g.* Finding a local minimum in a smooth function is NP-Hard. Finding a 2nd order optimal point in a non-smooth function is NP-Hard.
- Follow the Gradient Principle: No known convergence results for even back-propagation to stationary points!

The State of Non-Convex Optimization

- Practical observation: Empirically, non-convexity is not an issue. Gradient methods find high quality solutions.
- Theoretical Side: NP-hard in many cases. *e.g.* Finding a local minimum in a smooth function is NP-Hard. Finding a 2nd order optimal point in a non-smooth function is NP-Hard.
- Follow the Gradient Principle: No known convergence results for even back-propagation to stationary points!

Question

1. Why is (stochastic) gradient descent (GD) successful? Or is it just "alchemy"?

Setting

(Sub)-Gradient Descent

Gradient Descent algorithm:

$$x_{k+1} = x_k - \alpha_k \partial f(x_k).$$

Non-smoothness

Deep Learning Loss Functions are not smooth! (e.g. ReLU, max-pooling, batch-norm)

Convergence of sub-gradient method to stationary points is only known for weakly-convex functions ($f(x) + \frac{\lambda}{2}\|x\|^2$ convex).

Non-smooth Non-convex Optimization

Theorem (Davis, Drusvyatskiy, Kakade, and Lee)

Let x_k be the iterates of the stochastic sub-gradient method. Assume that f is locally Lipschitz (and semialgebraic), then every limit point x^* is critical:

$$0 \in \partial f(x^*).$$

- Difficulty is in the downward "kinks" like $(1 - \text{ReLU}(x))^2$
- Convergence rate is polynomial in $\frac{1}{\epsilon}, d$ to ϵ-subgradient.
- Clarke subgradient can be efficiently computed using Automatic Differentiation in $6x$ cost as function evaluation (Kakade and Lee 2018)

Theorem (Lee et al., COLT 2016)

Let $f : \mathbf{R}^n \to \mathbf{R}$ be a twice continuously differentiable function with the strict saddle property, then gradient descent with a random initialization converges to a local minimizer or negative infinity.

- Theorem applies for many optimization algorithms including coordinate descent, mirror descent, manifold gradient descent, and ADMM (Lee et al. 2017 and Hong et al. 2018)
- Stochastic optimization with injected isotropic noise finds local minimizers in polynomial time (Pemantle 1992; Ge et al. 2015, Jin et al. 2017)

Why are local minimizers interesting?

All local minimizers are global for the following problems:

1. ReLU networks via landscape design (GLM18)
2. Matrix Completion (GLM16)
3. Rank k approximation
4. Matrix Sensing (BNS16)
5. Phase Retrieval (SQW16)
6. Orthogonal Tensor Decomposition (GHJY15)
7. Dictionary Learning (SQW15)
8. Max-cut via Burer Monteiro (BBV16, Montanari 16)
9. Overparametrized Deep Networks (DL18)

Over-parametrization

If back-propagation is not finding a low training error solution, then fit a bigger model.

Problem

How much over-parametrization do we need to efficiently optimize?

Previous Work

Over-parametrization Hypothesis

Optimization is "easy" when parameters $>$ sample size (specialized to two-layer nets).

- Soudry and Carmon 2016 justified this for ReLU networks.
- Livni et al. empirically demonstrated that over-parametrization is necessary for SGD to work.
- When $\#$ neurons $>$ sample size , then all local are global for *unregularized training loss*. Easy to find global min by training only output layer or the "radial" component of the hidden layer.

Why Quadratic Activation?

Case Study: Quadratic Activation Networks

$$f(x;W) = \sum_{i=1}^{k} \phi(w_i^T x),$$

where $\phi(z) = z^2$.
These can be formulated as matrix sensing with $\mathcal{X}_i = x_i x_i^T$.

Regularized Loss

$$\min_W \sum_i \ell(f(x_i;W), y_i) + \frac{\lambda}{2} \|W\|_F^2 .$$

How much Over-parametrization?

- For $k \geq d$ that all local are global; relies on $y = x^T W^T W x = x^T M x$ for $M = W^T W$ (Haeffele and Vidal, Bach, Burer-Monteiro)
- The result is independent of n, which is counter-intuitive. Can we get closer to $\#$ params $= kd > n$?

Smoothed Analysis

Random Regularization

$$L_C(W) = \sum_i \ell(f(x_i; W), y_i) + \frac{\lambda}{2} \|W\|_F^2 + \langle C, W^T W \rangle,$$

where C is random Gaussian $\mathcal{N}(0, \sigma^2)$.

Theorem

Let ℓ be a convex loss function, $\lambda > 0$, and $\sigma > 0$. If $k \geq \sqrt{2n}$, then almost surely all local min are global minima.

- Applies for arbitrarily small perturbation σ. By choosing σ small, we can closely approximate the solution of the unperturbed objective.
- Motivated by work on SDP (Burer & Monteiro, Boumal-Voroniski-Bandeira) which show that $k \geq \sqrt{2n}$ all non-degenerate local minima are global. Smoothing allows us to remove the non-degenerate local minima.
- Suprisingly, the same smoothing works even though our objective is not SDP-representable.

How about Generalization?

Generalization

The regularizer $\|W\|_F^2$ corresponds to $\left\|W^T W\right\|_*$. Small nuclear norm leads to generalization via standard Rademacher complexity bounds.

Corollary

Assume that $y = \sum_{i=1}^{k_0} \sigma(w_l^T x)$, *and* $x_i \sim \mathcal{N}(0, I)$. *Then for* $n \gtrsim \frac{dk_0^2}{\epsilon^2}$,

$$L_{te}(W) - L_{tr}(W) \leq \epsilon.$$

The sample complexity is independent of k, *the number of neurons.*

Conclusion

Quadratic Activation Network

1. Training Error: Over-parametrization makes the optimization easy, since all local are global.
2. Test Error: The generalization is not hurt by over-parametrization. The sample complexity only depends on k_0 , the number of effective neurons, and not k, the number of neurons in the model.

How do we show this for ReLU activations and deeper networks?

Margin Theory for Neural Nets

Large margin

Do we obtain large margin classifiers from cross-entropy loss?

Let $f(\Theta; x)$ be the prediction function of a positive-homogeneous neural network.

Regularized Loss

$$\ell(f(\Theta; x)) + \lambda \|\Theta\|.$$

Theorem (Wei, Lee, Liu, Ma 2018)

Assume the dataset is separable by the network by normalized margin γ. Then the attained normalized margin by minimizing cross-entropy loss $\gamma_\lambda \to \gamma$.

- Overparametrization improves the optimal normalized margin: in two-layer networks, the margin

$$\gamma_1 < \ldots < \gamma_{n-1} < \gamma_n = \gamma_{n+1} = \ldots = \gamma_\infty$$

Theorem (Very Informal, see Openreview)

For a two-layer network that is infinitely wide (or $\exp(d)$ wide), gradient descent with noise converges to a global minimum of the regularized training loss.

- Overparametrization helps gradient descent find solutions that generalize.

Can Overparametrized Networks Generalize?

- Modern networks are over-parametrized meaning $p \gg n$ ($\frac{p}{n} \in (10, 200)$).
- Over-parametrization allows SGD to drive the training error to 0. But shouldn't the test error be huge due to overfitting?

Experiment

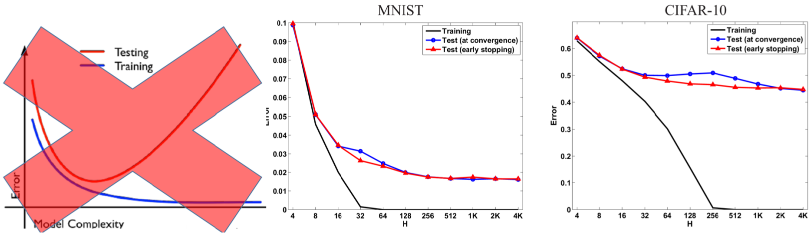

Figure: Credit: Neyshabur et al. See also Zhang et al.

- $p \gg n$, no regularization, no early stopping, and yet we do not overfit.
- Unclear what is the correct measure of model complexity. Clearly, parameter counting is not appropriate for SGD.

Experiment

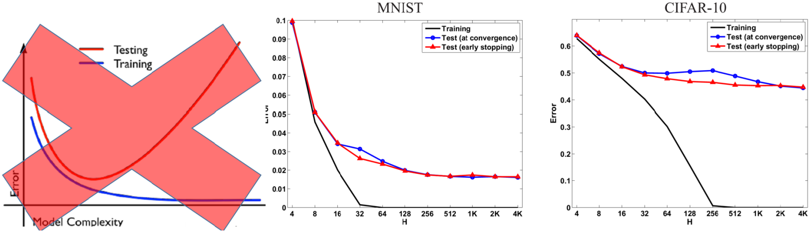

Figure: Credit: Neyshabur et al. See also Zhang et al.

- $p \gg n$, no regularization, no early stopping, and yet we do not overfit.
- Unclear what is the correct measure of model complexity. Clearly, parameter counting is not appropriate for SGD.
- Or is there regularization? Since $p \gg n$, there is a $p - n$-dimensional space of global minima, and definitely some of these do not generalize.

Today's Setting

Definition (Separable Data)

We will assume that $y_i(x_i^T w) > 0$ for some w.

- Equivalent of the over-parametrized regime in linear models. If $p \gg n$, this holds for almost all $\{x_i\}$.
- When the data is separable, there are infinitely many linear separators.

Implicit Regularization (via choice of Algorithm)

Warm-up: Logistic Regression with separable data

Gradient descent with any initial point w_0 on

$$\mathcal{L}(w) = \sum_i \log(1 + \exp(-y_i x_i^T w))$$

converges in direction to the ℓ_2-SVM solution. In equations,

$$\frac{w(t)}{\|w(t)\|} \to C \operatorname*{arg\,min}_{y_i w^T x_i \geq 1} \|w\|_2 .$$

(Soudry et al. 2018, Ji & Telgarsky 2018, Gunasekar et al. 2018)

This means that if the data is separable with a large margin, then GD+Logistic Regression generalizes as well as SVM.

Steepest Descent

$$w(t+1) = w(t) + \alpha \Delta w(t)$$
$$\Delta w(t) = \arg\min_{\|v\| \le 1} v^T \nabla L(w(t)).$$

Coordinate descent is steepest descent wrt $\|\cdot\|_1$ and signed gradient method is steepest descent wrt $\|\cdot\|_\infty$.

Theorem (Gunasekar, Lee, Soudry, and Srebro)

On separable data, steepest descent converges in direction to the $\|\cdot\|$-SVM solution, meaning $\frac{w(t)}{\|w(t)\|} \to C \arg\min_{y_i w^T x_i \geq 1} \|w\|$.

- Solution depends on the choice of algorithm.
- For coordinate descent, it is already known from the boosting literature that AdaBoost achieves the minimum ℓ_1 norm solution (Ratsch et al. 2004, Zhang and Yu 2005, Telgarsky 2013). Also related to the study of LARS algorithms.
- For ℓ_2 norm, this recovers the theorem before.

Theorem (Gunasekar, Lee, Soudry and Srebro 2018)

For any homogeneous polynomial p, GD on

$$\sum_i \exp(-y_i \langle p(w), \mathcal{X}_i \rangle)$$

converges to a first-order stationary point of

$$\min \|w\|_2 \quad \text{st } \langle p(w), \mathcal{X} \rangle \geq 1$$

Summary: Implicit Bias

Implicit Regularization

1. Overparametrize to make training easy, but there are infinitely many possible global minimum
2. The choice of algorithm and parametrization determine the global minimum.
3. Generalization is possible in the over-parametrized regime with no regularization by choosing the right algorithm.
4. We understand only very simple problems and algorithms.

References

Acknowledgements: This is joint work with the following co-authors below.

1. Wei, Lee, Liu, and Ma, *On the Margin Theory of Neural Networks.*
2. Gunasekar, Lee, Soudry and Srebro, *Characterizing Implicit Bias in Terms of Optimization Geometry.*
3. Du and Lee, *On the Power of Over-parametrization in Neural Networks with Quadratic Activation*
4. Davis, Drusvyatskiy, Sham Kakade, and Jason D. Lee, *Stochastic subgradient method converges on tame functions.*
5. Lee, Panageas, Piliouras, Simchowitz, Jordan, and Recht, *First-order Methods Almost Always Avoid Saddle Points.*
6. Lee, Simchowitz, Jordan, and Recht, *Gradient Descent Converges to Minimizers.*