Optimization geometry and implicit regularization

Suriya Gunasekar

Joint work with N. Srebro (TTIC), J. Lee (USC), D. Soudry (Technion), M.S. Nacson (Technion), B. Woodworth (TTIC), S. Bhojanapalli (TTIC), B. Neyshabur (TTIC-> IAS)
Optimization in ML

$h_W: x \rightarrow y$ parameterized by $W \in \mathbb{R}^d$

Training data $\{(x_n, y_n): n = 1, 2, \ldots, N\}$

\[
\hat{W} = \arg\min_W \sum_{n=1}^{N} \ell(h_W(x_n), y_n)
\]
Optimization in ML

\( h_w : x \rightarrow y \) parameterized by \( W \in \mathbb{R}^d \)

Training data \( \{(x_n, y_n) : n = 1, 2, \ldots, N\} \)

\[ \hat{W} = \arg\min_W \sum_{n=1}^{N} \ell(h_w(x_n), y_n) \]

- Over parameterization: \( d \gg N \)
Optimization in ML

$h_w: x \rightarrow y$ parameterized by $W \in \mathbb{R}^d$

Training data $\{(x_n, y_n): n = 1, 2, \ldots, N\}$

$$\hat{W} = \arg\min_W \sum_{n=1}^N \ell(h_W(x_n), y_n)$$

- Over parameterization: $d \gg N$
- Many global minima – all have $\sum_{n=1}^N \ell(h_{\hat{W}}(x_n), y_n) = 0$
Optimization in ML

\( h_W: x \rightarrow y \) parameterized by \( W \in \mathbb{R}^d \)

Training data \( \{(x_n, y_n): n = 1, 2, \ldots N\} \)

\[
\hat{W} = \arg\min_W \sum_{n=1}^{N} \ell(h_W(x_n), y_n)
\]

- Over parameterization: \( d \gg N \)
- Many global minima – all have \( \sum_{n=1}^{N} \ell(h_{\hat{W}}(x_n), y_n) = 0 \)
- What we really care about is \( \mathbb{E}_{x,y} \ell(h_{\hat{W}}(x), y) \)
  \( \rightarrow \) Different global optima have different \( \mathbb{E}_{x,y} \ell(h_{\hat{W}}(x), y) \)
Learning overparameterized models

\[ h_W: x \rightarrow y \text{ parameterized by } W \in \mathbb{R}^d \]

\[ \hat{W} = \arg\min_W \sum_{n=1}^{N} \ell(h_W(x_n), y_n) + \mathcal{R}(W) \]

small \( \mathcal{R}(\hat{W}) \Rightarrow \)
small \( \mathbb{E}_{x,y} \ell(h_{\hat{W}}(x), y) - \frac{1}{n} \sum_{n=1}^{N} \ell(h_{\hat{W}}(x_n), y_n) \)

Explicit regularization for high dimensional estimation
Learning overparameterized models

\[ h_W : x \rightarrow y \text{ parameterized by } W \in \mathbb{R}^d \]

\[ \hat{W} = \arg\min_W \sum_{n=1}^{N} \ell(h_W(x_n), y_n) + \mathcal{R}(W) \]

\[ N \ll d \]

What happens if we don’t have \( \mathcal{R}(W) \)?
Matrix Estimation from Linear Measurements

\[
\min_{W \in \mathbb{R}^{d \times d}} L(W) := \sum_{n=1}^{N} (\langle X_n, W \rangle - y_n)^2 := \| \mathcal{X}(W) - y \|^2_2
\]

e.g. matrix completion, linear neural networks,…

➢ When \( N \ll d^2 \) optimization is underdetermined with many trivial global minima

e.g. impute 0 or 42 or 1321234123 for matrix completion
Matrix Estimation from Linear Measurements

\[
\min_{W \in \mathbb{R}^{d \times d}} L(W) := \sum_{n=1}^{N} (\langle X_n, W \rangle - y_n)^2 := \| \mathcal{X}(W) - y \|_2^2
\]

e.g matrix completion, linear neural networks,…

➤ When \( N \ll d^2 \) optimization is underdetermined with many trivial global minima

e.g. impute 0 or 42 or 1321234123 for matrix completion

\[
\min_{U, V \in \mathbb{R}^{d \times d}} \tilde{L}(U, V) = L(UV^\top) = \| \mathcal{X}(UV^\top) - y \|_2^2
\]

No explicit regularization & no rank constraint

➤ same trivial global minima exists
Matrix Estimation from Linear Measurements

\[
\min_{W \in \mathbb{R}^{d \times d}} L(W) := \sum_{n=1}^{N} (\langle X_n, W \rangle - y_n)^2 := \| \mathcal{X}(W) - y \|_2^2
\]

e.g. matrix completion, linear neural networks,…

➢ When \( N \ll d^2 \) optimization is underdetermined with

many trivial global minima

e.g. impute 0 or 42 or 1321234123 for matrix completion

\[
\min_{U, V \in \mathbb{R}^{d \times d}} \tilde{L}(U, V) = L(UV^\top) = \| \mathcal{X}(UV^\top) - y \|_2^2
\]

No explicit regularization & no rank constraint

➢ same trivial global minima exists

\[
U_{k+1} = U_k - \eta \nabla_U \tilde{L}(U_k, V_k)
\]

\[
V_{k+1} = V_k - \eta \nabla_V \tilde{L}(U_k, V_k)
\]

Gradient descent on \( \tilde{L}(U, V) \)
Gradient descent on $L(U)$ gets to “good” global minima
Gradient descent on $\tilde{L}(U)$ gets to “good” global minima

Gradient descent on $\tilde{L}(U)$ generalizes better with smaller step size
Question: Which global minima does gradient descent reach? Why does it generalize well?
Implicit Regularization

Different optimization algorithms

⇒ different global minimum $\hat{W}$

⇒ different generalization $E_{x,y}\ell(h_{\hat{W}}(x), y)$
Overparameterization in neural networks

- Image datasets
  - CIFAR ~ 60K images,
  - ImageNet ~14M images, ~1M annotations
- Architectures for vision tasks:
  - AlexNet (2012): 8 layers, 60M parameters
  - VGG-16 (2014): 16 layers, 138M parameters

NNs trained using local search have good generalization even without explicit regularization or early stopping

Bias of optimization algorithms

- Effect of optimization geometry (Neyshabur et al. 2015)
Bias of optimization algorithms

- Effect of optimization geometry (Neyshabur et al. 2015)
Bias of optimization algorithms

- Effect of optimization geometry (Neyshabur et al. 2015)
- Effect of size of minibatch (Keskar et al. 2017, Dihn et al. 2017)
- Effect of adaptive algorithms (Wilson et al. 2017)
- Learning to learn (Abdrychowicz et al. 2016, Finn et al. 2017)
Implicit Regularization

Different optimization algorithms
⇒ different global minimum $\hat{W}$
⇒ different generalization $\mathbb{E}_{x,y} \ell(h_{\hat{W}}(x), y)$

Can we characterize which specific global minimum different optimization algorithms converge to?
Implicit Regularization

Can we characterize which *specific* global minimum different optimization algorithms converge to?

How does this depend on optimization geometry, initialization, step size, momentum, stochasticity?
Implicit Regularization

Can we characterize which *specific* global minimum different optimization algorithms converge to?

How does this depend on optimization geometry, initialization, step size, momentum, stochasticity?

Understanding the implicit bias could enable

- Optimization algorithms for faster convergence AND better generalization
- New regularization techniques
- Efficiently train smaller networks
Gradient descent: linear regression

\[
\min_w L(w) = \sum_{n=1}^{N} (\langle x_n, w \rangle - y_n)^2
\]
Gradient descent: linear regression

$$\min_w L(w) = \sum_{n=1}^{N} (\langle x_n, w \rangle - y_n)^2$$

Gradient descent initialized at $w(0)$

$$w(t + 1) = w(t) - \eta \nabla_w L(w(t))$$

$$\nabla_w L(w(t)) = \sum_{n=1}^{N} (\langle x_n, w(t) \rangle - y_n) x_n$$
Gradient descent: linear regression

\[
\min_w L(w) = \sum_{n=1}^{N} (\langle x_n, w \rangle - y_n)^2
\]

Gradient descent initialized at \(w(0)\)

\[
w(t + 1) = w(t) - \eta \nabla_w L(w(t))
\]

\[
\nabla_w L(w(t)) = \sum_{n=1}^{N} (\langle x_n, w(t) \rangle - y_n)x_n
\]

Updates lie on a **low dimensional** affine manifold \(\Delta w(t) \in \text{span}(x_n)\)
Gradient descent: linear regression

$$\min_w L(w) = \sum_{n=1}^{N} (\langle x_n, w \rangle - y_n)^2$$

Gradient descent initialized at $w(0)$
$$w(t + 1) = w(t) - \eta \nabla_w L(w(t))$$

$$\nabla_w L(w(t)) = \sum_{n=1}^{N} (\langle x_n, w(t) \rangle - y_n)x_n$$

Updates lie on a **low dimensional** affine manifold $\Delta w(t) \in \text{span}(x_n)$

If $w(0) = 0$
$$w(t) \to \text{argmin}_{Xw=y} \|w\|_2$$
Gradient descent: linear regression

\[
\min_w L(w) = \sum_{n=1}^{N} (\langle x_n, w \rangle - y_n)^2
\]

Gradient descent initialized at \(w(0)\)

\[
w(t + 1) = w(t) - \eta \nabla_w L(w(t))
\]

\[
\nabla_w L(w(t)) = \sum_{n=1}^{N} (\langle x_n, w(t) \rangle - y_n) x_n
\]

Updates lie on a low dimensional affine manifold \(\Delta w(t) \in \text{span}(x_n)\)

If \(w(0) = 0\)

\[
w(t) \rightarrow \arg\min_{Xw=y} \|w\|_2
\]

\[
w(t) \rightarrow \arg\min_{Xw=y} \|w - w(0)\|_2
\]
Gradient descent: linear regression

\[
\min_w L(w) = \sum_{n=1}^{N} (\langle x_n, w \rangle - y_n)^2
\]

Gradient descent initialized at \(w(0)\)

\[
w(t + 1) = w(t) - \eta \nabla_w L(w(t))
\]

\[
\nabla_w L(w(t)) = \sum_{n=1}^{N} (\langle x_n, w(t) \rangle - y_n)x_n
\]

Updates lie on a low dimensional affine manifold \(\Delta w(t) \in \text{span}(x_n)\)

- If \(w(0) = 0\)
  \[w(t) \rightarrow \arg\min_{Xw=y} \|w\|_2\]

- \[w(t) \rightarrow \arg\min_{Xw=y} \|w - w(0)\|_2\]

Independent of step size \(\eta\), momentum, instancewise stochastic gradient descent
Gradient descent: linear regression

\[ \min_w L(w) = \sum_{n=1}^{N} (\langle x_n, w \rangle - y_n)^2 \]

Gradient descent initialized at \( w(0) \)

\[ w(t + 1) = w(t) - \eta \nabla_w L(w(t)) \]
\[ \nabla_w L(w(t)) = \sum_{n=1}^{N} (\langle x_n, w(t) \rangle - y_n) x_n \]

Updates lie on a **low dimensional** affine manifold
\[ \Delta w(t) \in \text{span}(x_n) \]
\[ \hat{w}(s)_{gd} = \arg \min_{Xw=y} \|w - w(0)\|_2 \]

**Same argument for linear models** \( \hat{y}(x) = \langle w, x \rangle \) and **loss functions** \( \ell(\hat{y}(x), y) \) with unique finite root at \( \hat{y} = y \)
Can we get such results for other problems and other optimization algorithms?

First, same problem different optimization algorithms
Mirror descent w.r.t potential $\psi$

$$w(t + 1) = \arg\min_{w} \eta \langle w, \nabla_w L(w(t)) \rangle + \frac{1}{2} \|w - w(t)\|_2^2$$
Mirror descent w.r.t potential $\psi$

Gradient descent

$$w(t + 1) = \operatorname{argmin}_w \eta \langle w, \nabla_w L(w(t)) \rangle + \frac{1}{2} \| w - w(t) \|_2^2$$

Mirror Descent w.r.t. strongly convex potential $\psi$

$$w(t + 1) = \operatorname{argmin}_w \eta \langle w, \nabla_w L(w(t)) \rangle + D_\psi (w, w(t))$$

$$D_\psi (w, w(t)) = \psi(w) - \psi(w(t)) - \langle \nabla \psi(w(t)), w - w(t) \rangle$$

e.g. $\psi(w) = \sum_i w[i] \log w[i] \Rightarrow D_\psi (w, w(t)) = KL(w, w(t))$
Mirror descent w.r.t potential $\psi$

$$w(t + 1) = \arg \min_w \eta \langle w, \nabla_w L(w(t)) + D_\psi (w, w(t)) \rangle$$

$$\nabla \psi(w(t + 1)) = \nabla \psi(w(t)) - \sum_{n=1}^{N} \ell'(w(t))x_n$$
Mirror descent w.r.t potential $\psi$

$$w(t + 1) = \arg\min_w \eta \langle w, \nabla_w L(w(t)) + D_\psi (w, w(t)) \rangle$$

$$\nabla \psi(w(t + 1)) = \nabla \psi(w(t)) - \sum_{n=1}^{N} \ell'(w(t))x_n$$

Dual updates lie on the low dimensional affine manifold $\Delta w(t) \in \text{span}(x_n)$

If $\nabla \psi(w(0)) = 0$

$$w(t) \rightarrow \arg\min_{Xw=y} \psi(w)$$

$$w(t) \rightarrow \arg\min_{Xw=y} D_\psi (w, w(0))$$

G, Lee, Soudry, Srebro. Arxiv 2018
Mirror descent w.r.t potential $\psi$

$$w(t + 1) = \arg\min_w \eta \langle w, \nabla_w L(w(t)) \rangle + D_\psi (w, w(t))$$

$$\nabla \psi(w(t + 1)) = \nabla \psi(w(t)) - \sum_{n=1}^{N} \ell'(w(t)) x_n$$

Dual updates lie on the low dimensional affine manifold $\Delta w(t) \in \text{span}(x_n)$

If $\nabla \psi(w(0)) = 0$

\[ w(t) \rightarrow \arg\min_{Xw=y} \psi(w) \]

\[ \rightarrow \arg\min_{Xw=y} D_\psi (w, w(0)) \]

- Again independent of step size, stochasticity, dual momentum
- Also works with affine constraints on $w$

Exponentiated gradient descent

$\Rightarrow$ implicit entropic regularization $\psi(w) = \sum_i w[i] \log(w[i])$

G, Lee, Soudry, Srebro. Arxiv 2018
**Steepest descent w.r.t. norm $\| \cdot \|_*$**

Gradient descent

$$w(t + 1) = w(t) + \eta \Delta w(t)$$

$$\Delta w(t) = \arg\min_{v: \|v\|_2 \leq 1} \langle v, \nabla_w L(w(t)) \rangle$$

Steepest Descent w.r.t. general norm $\| \cdot \|$

$$w(t + 1) = w(t) + \eta \Delta w(t)$$

$$\Delta w(t) = \arg\min_{v: \|v\| \leq 1} \langle v, \nabla_w L(w(t)) \rangle$$

e.g. Coordinate descent $\| \cdot \| = \| \cdot \|_1$
Steepest descent w.r.t. norm \( \| \cdot \| \)

Gradient descent

\[
w(t + 1) = w(t) + \eta \Delta w(t)
\]

\[
\Delta w(t) = \arg\min_{v: \|v\|_2 \leq 1} \langle v, \nabla_w L(w(t)) \rangle
\]

Steepest Descent w.r.t. general norm \( \| \cdot \| \)

\[
w(t + 1) = w(t) + \eta \Delta w(t)
\]

\[
\Delta w(t) = \arg\min_{v: \|v\| \leq 1} \langle v, \nabla_w L(w(t)) \rangle
\]

e.g. Coordinate descent \( \| \cdot \| = \| \cdot \|_1 \)

\[
w(t) \rightarrow \arg\min_{Xw=y} \|w - w(0)\|
\]
Steepest descent w.r.t. norm $\|\cdot\|$ 

$$w(t + 1) = w(t) + \eta \Delta w(t)$$

$$\Delta w(t) = \arg\min_{v: \|v\| \leq 1} \langle v, \nabla_w L(w(t)) \rangle$$

Even for $\eta \to 0$: 
$$w(t) \nleftrightarrow \arg\min_{Xw=y} \|w - w(0)\|$$

G, Lee, Soudry, Srebro. Arxiv 2018
Can we get such results for other problems and other optimization algorithms?

How about gradient descent on other problems or different parameterizations?
Matrix Estimation from Linear Measurements

\[
\min_{W \in \mathbb{R}^{d \times d}} L(W) := \sum_{n=1}^{N} (\langle X_n, W \rangle - y_n)^2 := \| \mathcal{X}(W) - y \|^2_2
\]

e.g. matrix completion, linear neural networks,…

➢ When \( N \ll d^2 \) optimization is underdetermined with many trivial global minima

e.g. impute 0 or 42 or 1321234123 for matrix completion

\[
\min_{U, V \in \mathbb{R}^{d \times d}} \tilde{L}(U, V) = L(UV^\top) = \| \mathcal{X}(UV^\top) - y \|^2_2
\]

No explicit regularization & no rank constraint

➢ same trivial global minima exists

Gradient descent on \( \tilde{L}(U, V) \)

\[
U_{k+1} = U_k - \eta \nabla_U \tilde{L}(U_k, V_k)
\]
\[
V_{k+1} = V_k - \eta \nabla_V \tilde{L}(U_k, V_k)
\]
Question: Which global minima does gradient descent reach? Why does it generalize well?

\[ d = 50, \ N = 300, \ X_n \text{ i.i.d Gaussian}, \ W^* \text{ rank-2 ground truth} \]
\[ y = X(W^*) + N(0, 10^{-3}), \ y_{\text{test}} = X_{\text{test}}(W^*) + N(0, 10^{-3}) \]
Gradient descent on $\tilde{L}(U)$ converges to a minimum nuclear norm solution.
Gradient descent on $\tilde{L}(U, V)$ converges to the minimum nuclear norm solution

$$W(t) = U(t)V(t)^\top \rightarrow W_{NN}^* = \arg\min_{W} \|W\|_*^{\star} \quad \mathcal{X}(W) = y$$

when,

- Initialization is close to 0
- Step size is very small \( \dot{U}_t = \frac{dU_t}{dt} = -\nabla_U \tilde{L}(U) \)
Commutative $X_i$

\[ X_iX_j = X_jX_i \text{ for all } i, j \in [N] \]

\[ W(t) = e^{\lambda^*(s_t)} W(0) e^{-\lambda^*(s_t)} \text{ for some } s_t \in \mathbb{R}^N \]

$\eta \rightarrow 0$ necessary to remain in the (non-linear) manifold
Commutative $X_i$

$X_iX_j = X_jX_i$ for all $i, j \in [N]$

$W(t) = e^{x^*(s_t)} W(0) e^{-x^*(s_t)}$ for some $s_t \in \mathbb{R}^N$

$\eta \to 0$ necessary to remain in the (non-linear) manifold

Let $U_\infty(\alpha)$ be the solution of gradient flow initialized at $U_0 = \alpha I$.

If measurements $X_n$ commute, i.e. $X_iX_j = X_jX_i$, and

if $\bar{W}_\infty = \lim_{\alpha \to 0} U_\infty(\alpha) U_\infty(\alpha)^\top$ exists and satisfies $L(\bar{W}_\infty) = 0$, then

$$\bar{W}_\infty = W_{\text{NN}}^* = \min_{\mathcal{X}(W)=y} \|W\|_*$$

Conjecture proved for RIP $X_n$ by Li et al. (2018)
Proof Ideas

- Characterize the manifold in which the $w(t)$ lie on
  - $w(0) + \text{span}(x_n)$ for gradient descent
  - $\nabla \psi^{-1} \left( \nabla \psi(w(0)) + \text{span}(x_n) \right)$ for mirror descent
  - $e^{x^*(s)}W(0)e^{x^*(s)}$ for matrix factorization with
    $\eta \to 0, \|W(0)\| \to 0$, commutative $X_n$
Proof Ideas

• Characterize the manifold in which the $w(t)$ lie on
  $\rightarrow \ w(0) + \text{span}(x_n)$ for gradient descent
  $\rightarrow \nabla \psi^{-1} \left( \nabla \psi(w(0)) + \text{span}(x_n) \right)$ for mirror descent
  $\rightarrow \ e^{x^*(s)}W(0)e^{x^*(s)}$ for matrix factorization with
    $\eta \rightarrow 0, \|W(0)\| \rightarrow 0$, commutative $X_n$

• Show that all the global minima on the manifold satisfy the KKT conditions for “regularized” problem
  $\rightarrow \ \min_{XW=y} \|w - w(0)\|_2$
  $\rightarrow \ \min_{XW=y} D_\psi(w, w(0))$
  $\rightarrow \ \min_{X(W)=y} \|W\|_*$
Losses with a unique finite root

- Robust characterization of for general mirror descent with potential $\psi$
  \[ w(t) \to \min_{xw=y} D_\psi(w, w(0)) \]

- No useful characterization for generic steepest descent w.r.t norm $\|\cdot\|$
  $\to$ even when $\|\cdot\|^2$ strongly convex
  $\to$ even for $\eta \to 0$

- Fragile characterization for matrix factorization
  \[ W(t) \to \min_{W \succeq 0, x(W) = y} \|W\|_* \]
  $\to$ ONLY for $\|W(0)\| \to 0, \eta \to 0$
  $\to$ Proven only for RIP measurements
  $\to$ initialization close to 0 is particularly bad!!
Losses with a unique finite root

- Robust characterization of for general mirror descent with potential $\psi$
  $$w(t) \to \min_{xw=y} D_\psi(w, w(0))$$

- No useful characterization for generic steepest descent w.r.t. norm $\|\|$
  $\Rightarrow$ even when $\|\|^2$ strongly convex
  $\Rightarrow$ even for $\eta \to 0$

- Fragile characterization for matrix factorization
  $$W(t) \to \min_{W \succeq 0, x(W)=y} \|W\|_*$$
  $\Rightarrow$ ONLY for $\|W(0)\| \to 0$, $\eta \to 0$
  $\Rightarrow$ Proven only for RIP measurements
  $\Rightarrow$ initialization close to 0 is particularly bad!!

What happens with other losses?
Losses with a unique finite root

- Robust characterization of for general mirror descent with potential $\psi$
  $$w(t) \to \min_{x \in \{y\}} D_\psi(w, w(0))$$

- No useful characterization for generic steepest descent w.r.t norm $\|\cdot\|$ even when $\|\cdot\|^2$ strongly convex even for $\eta \to 0$

- Fragile characterization for matrix factorization
  $$W(t) \to \min_{W \geq 0, x(W) = y} \|W\|_*$$
  ONLY for $\|W(0)\| \to 0, \eta \to 0$
  Proven only for RIP measurements
  initialization close to 0 is particularly bad!!

What happens with other losses?
Very different for logistic regression – no finite minima
Implicit bias when global minimum is unattainable!

Logistic regression on separable data
Gradient descent: logistic regression

\[
\min_w L(w) = \sum_{n=1}^{N} \log(1 + \exp(-y_n \langle x_n, w \rangle))
\]

Gradient descent initialized at \(w(0)\)
\[
w(t + 1) = w(t) - \eta \nabla_w L(w(t))
\]

\[
\nabla_w L(w(t)) = \sum_{n=1}^{N} r_n(t) x_n
\]

Soudry, Hoffer, Srebro, ICLR 2018
Gradient descent: logistic regression

\[
\min_w L(w) = \sum_{n=1}^{N} \log(1 + \exp(-y_n \langle x_n, w \rangle))
\]

Gradient descent initialized at \( w(0) \)

\[
w(t + 1) = w(t) - \eta \nabla_w L(w(t))
\]

\[
\nabla_w L(w(t)) = \sum_{n=1}^{N} r_n(t) x_n
\]

but \( \|w(t)\| \to \infty \)
Gradient descent: logistic regression

\[
\min_w L(w) = \sum_{n=1}^{N} \log(1 + \exp(-y_n \langle x_n, w \rangle))
\]

Gradient descent initialized at \( w(0) \)

\[
w(t + 1) = w(t) - \eta \nabla_w L(w(t))
\]

\[
\nabla_w L(w(t)) = \sum_{n=1}^{N} r_n(t) x_n
\]

but \( \|w(t)\| \to \infty! \)

\[
\frac{w(t)}{\|w(t)\|_2} \to \arg\max \min_w \sum_n y_n \langle w, x_n \rangle \\
\text{subject to } \|w\|_2 \leq 1
\]

Independent of step size \( \eta \) and initialization \( w(0) \)

Soudry, Hoffer, Srebro, ICLR 2018
Gradient descent: logistic regression

\[
\min_w L(w) = \sum_{n=1}^{N} \log(1 + \exp(-y_n \langle x_n, w \rangle))
\]

Gradient descent initialized at \(w(0)\)
\[
w(t + 1) = w(t) - \eta \nabla_w L(w(t))
\]

\[
\nabla_w L(w(t)) = \sum_{n=1}^{N} r_n(t) x_n
\]

but \(\|w(t)\| \to \infty!\)
\[
\frac{w(t)}{\|w(t)\|_2} \to \arg \max_{w: \|w\|_2 \leq 1} \min_n y_n \langle w, x_n \rangle
\]

Holds for linear classifiers \(\hat{y}(x) = \langle w, x \rangle\) and any strictly monotone loss \(\ell(\hat{y}(x), y)\) with exponential tail

Soudry, Hoffer, Srebro, ICLR 2018
How fast is the margin maximized?

Fixed step size $\eta$

- $O\left(\frac{1}{\log(t)}\right)$ - extremely slow!!
- Compare with $O\left(\frac{1}{t}\right)$ convergence of $L(w)$

---

Schpiegel, Lee, G, Srebro, Soudry, Arxiv 2018
How fast is the margin maximized?

Fixed step size $\eta$
- $O\left(\frac{1}{\log(t)}\right)$ - extremely slow!!
- Compare with $O\left(\frac{1}{t}\right)$ convergence of $L(w)$

Can we use lighter or heavier tail to get faster convergence?
No. exponential-tail yields optimal rate.
- For $\ell(u) = \exp(-u^v), v > 1$, margin converges as $O\left(\frac{1}{\log^{1/v} t}\right)$
- For $\ell(u) = \exp(-u^v), \frac{1}{4} \leq v < 1$, margin converges as $\frac{c}{v \log t}$
- For $\ell(u) \propto u^{-v}$, does not converge to max-margin

Soudry, Hoffer, Srebro, 2018
How fast is the margin maximized?

Fixed step size $\eta$

- $O\left(\frac{1}{\log(t)}\right)$ - extremely slow!!
- Compare with $O\left(\frac{1}{t}\right)$ convergence of $L(w)$

Can we use lighter or heavier tail to get faster convergence?

No. exponential-tail yields optimal rate.

- For $\ell(u) = \exp(-u^v), v > 1$, margin converges as $O\left(\frac{1}{\log^{1/v} t}\right)$
- For $\ell(u) = \exp(-u^v), \frac{1}{4} \leq v < 1$, margin converges as $\frac{c}{v \log t}$
- For $\ell(u) \propto u^{-v}$, does not converge to max-margin

Any other way to get faster convergence?

Yes. Stepsize $\eta \propto 1/\|\nabla L\|$ yield $\tilde{O}\left(\frac{1}{\sqrt{t}}\right)$ convergence (comparable to best hard-margin SVM algorithms)
Implicit bias on strictly monotone losses with exponential tail

Can we get a more robust characterization compared to regression-type losses?
Steepest descent w.r.t. norm $\| \cdot \|$ 

$$w(t + 1) = w(t) + \eta \Delta w(t)$$

$$\Delta w(t) = \arg\min_{v: \|v\| \leq 1} \langle v, \nabla_w L(w(t)) \rangle$$
Steepest descent w.r.t. norm $\|\cdot\|$ 

$$w(t + 1) = w(t) + \eta \Delta w(t)$$

$$\Delta w(t) = \arg\min_{v: \|v\| \leq 1} \langle v, \nabla_w L(w(t)) \rangle$$

$$\frac{w(t)}{\|w(t)\|} \rightarrow \max_{w: \|w\| \leq 1} \min_n y_n \langle w, x_n \rangle$$

$\rightarrow$ Independent of initialization
$\rightarrow$ Small enough $\eta$
Steepest descent w.r.t. norm $\|\cdot\|$:

$w(t + 1) = w(t) + \eta \Delta w(t)$

$\Delta w(t) = \arg\min_{v : \|v\| \leq 1} \langle v, \nabla_w L(w(t)) \rangle$

$\frac{w(t)}{\|w(t)\|} \rightarrow \max_{w : \|w\| \leq 1} \min_n y_n \langle w, x_n \rangle$

→ Independent of initialization
→ Small enough $\eta$

Compare with squared loss →

G, Lee, Soudry, Srebro. Arxiv 2018
Matrix Estimation from Linear Measurements

e.g. matrix completion, linear neural networks,…

➢ When $N \ll d^2$ optimization is underdetermined with many trivial global minima
e.g. impute 0 or 42 or 1321234123 for matrix completion

Gradient descent on $\tilde{L}(U, V)$

$$U_{k+1} = U_k - \eta \nabla_U \tilde{L}(U_k, V_k)$$
$$V_{k+1} = V_k - \eta \nabla_V \tilde{L}(U_k, V_k)$$
Matrix Estimation from Linear Measurements

- e.g. matrix completion, linear neural networks, ...

- When $N \ll d^2$ optimization is underdetermined with many trivial global minima
  
  - e.g. impute 0 or 42 or 1321234123 for matrix completion

Gradient descent on $\tilde{L}(U, V)$

$$
U_{k+1} = U_k - \eta \nabla_U \tilde{L}(U_k, V_k) \\
V_{k+1} = V_k - \eta \nabla_V \tilde{L}(U_k, V_k)
$$

Let $W(t) = U(t)U(t)^\top$. For any full rank $W(0)$ and any $\eta_t$ such that $\{L(W(t))\}_t$ is strictly decreasing, if $\frac{\Delta W(t)}{||\Delta W(t)||}$ and $\frac{\nabla L(W(t))}{||\nabla L(W(t))||}$ exists, then

$$
\frac{W(t)}{||W(t)||_*} \rightarrow \max_{||W||_* \leq 1} \min_n y_n \langle X_n, W \rangle
$$

G, Lee, Soudry, Srebro. Arxiv 2018
Strictly monotone losses

- **Gradient descent**
  \[
  \frac{w(t)}{\|w(t)\|_2} \rightarrow \max_{\|w\|_2 \leq 1} \min_n y_n \langle x_n, w \rangle
  \]
  
  → Independent of initialization
  
  → Any step size leading to descent algorithm

- **Steepest descent w.r.t norm \|\cdot\|**
  \[
  \frac{w(t)}{\|w(t)\|} \rightarrow \max_{\|w\| \leq 1} \min_n y_n \langle x_n, w \rangle
  \]
  
  → Independent of initialization
  
  → Any step size leading to descent algorithm

- **Matrix factorization**
  \[
  \frac{W(t)}{\|W(t)\|_*} \rightarrow \max_{\|W\|_* \leq 1} \min_n y_n \langle X_n, W \rangle
  \]
  
  → Independent of initialization
  
  → Any step size leading to descent algorithm
Simplicity from Asymptotics

Squared loss:
- \( w(\infty) \) depends on initial \( w(0) \) and stepsize \( \eta \)
- May need to take \( \eta \to 0, w(0) \to 0 \) to get characterization in terms of gradient manifold

Exponential loss
- \( \frac{w(\infty)}{\|w(\infty)\|} \) does NOT depend on initial \( w(0) \) and stepsize \( \eta \)
- What happens at the beginning doesn’t effect the asymptotic behavior as \( \|w(\infty)\| \to \infty \)
- Limit direction dominated only by the updates and hence the gradient manifold
• Role of optimization in ML extends beyond reaching some global minima
• Implicit regularization plays a crucial role in generalization of over parameterized models
• Understanding specific global minimum reached by an algorithm is important!