Implicit Regularization

Last time: Gradient descent on a wide-enough deep net reaches a zero error solution

But there are often tons of zero error solutions!

Main Question: which zero error soln. do we reach? And how do reparameterizations change things?

We'll study underdetermined least squares:

$$\ell(x) = \frac{\|Ax - b\|_2^2}{2}$$

where A is $n \times m$ and $n < m$

Consider the gradient flow:

$$x(0) = 0 \, , \; \dot{x}(t) = -\nabla \ell(x(t))$$

Lemma 1: $x(t) \longrightarrow \underset{Ax=b}{\operatorname{argmin}} \|x\|_2$

Proof: (sketch) A priori it is not clear that $x(t)$ converges to a zero error solution.

We will take this for granted

It's related to standard convergence for GD

Let x^* be the point it converges to

Next we claim $x^* \in \operatorname{span}(A^T)$

The stronger statement that

$$x(t) \in \operatorname{span}(A^T)$$

follows because

$$x(t) = x(t) - x(0)$$

$$= \int_0^t -\nabla \ell(x(s))\, ds$$

$$= \int_0^t -A^T(Ax(s) - b)\, ds$$

Finally,

$$x^* \in \text{span}(A^T) \Rightarrow x^* \perp \ker(A)$$

$$\Rightarrow x^* = \underset{Ax=b}{\text{argmin}}\ \|x\|_2 \qquad \blacksquare$$

Thus the point we reach is the soln. to the following regularized problem

$$\min_x \frac{\|Ax - b\|_2^2}{2} + \lambda \|x\|_2^2$$

in the limit as $\lambda \to 0$

Today we will explore what happens when we take our original least squares problem and reparameterize it

def: Let sq(y) denote the entrywise square

Now consider the reparameterized objective:

$$\hat{\ell}(y) = \frac{\| A\, sq(y) - b \|_2^2}{2}$$

If we let x = sq(y) we are restricting to nonnegative x

But something more interesting happens in terms of which zero error soln (with nonneg. entries) we reach

In particular consider:

$$y(0) = \alpha \vec{1} \quad \text{and} \quad \dot{y}(t) = -\nabla \ell(y(t))$$

and set $x(t) \triangleq sq(y(t))$

Main Theorem: Suppose $x(t)$ converges to a zero error soln $x^*(\alpha)$. Then as $\alpha \to 0$ we have

$$x^*(\alpha) \longrightarrow \underset{x \in \mathbb{R}^n_{\geq 0},\, Ax=b}{\text{argmin}} \|x\|_1$$

Thus by reparameterizing we've changed the mechanism by which we select a zero error soln.

Side remark: Solving the optimization problem

$$\text{(BP)} \quad \min \|x\|_1 \quad \text{s.t. } Ax=b$$

is useful in compressed sensing

Main theorem (informal) can recover a k-sparse n-dimensional x from

$$O(k \log \tfrac{n}{k})$$

random linear measurements and solving (BP) = "basis pursuit"

Next let's describe the trajectory in x-space:

$$\dot{x}(t) = 2 D_y \dot{y}(t), \quad D_y = \text{diag}(y(t))$$

Now we can compute

$$\dot{y}(t) = -\nabla \left(\frac{\|A\,\text{sq}(y(t)) - b\|_2^2}{2} \right)$$

$$= -2 D_y A^T (A\,\text{sq}(y(t)) - b)$$

Putting it all together:

$$\dot{x}(t) = -4 D_x A^T (A x(t) - b)$$

$D_x = \text{diag}(x(t))$

Informal Claim: The D_x term is changing the local geometry

i.e. the gradient points in the direction of steepest ascent, but if we change how we measure closeness, it will point in a different direction

Proof Outline

(1) study mirror descent, which is a standard way to adjust gradient updates to fit problem geometry

(2) Interpret the reparameterized gradient flow as mirror descent

Mirror Descent

First let's give a reformulation of vanilla GD:

<u>MD with Euclidean Distance</u>

For $t = 1$ to T

Set $x_{t+1} =$

$$\arg\min_{x} \Big\{ \underbrace{\cancel{f(x_t)} + \langle \nabla f(x_t), x - x_t \rangle}_{\text{linear approx at } x_t} + \frac{1}{2\eta} \|x - x_t\|_2^2 \Big\}$$

Let's check that this indeed recovers vanilla GD

Computing the gradient of the expression we want to minimize, and setting it equal to zero at $x = x_{t+1}$ we get:

$$\left[\nabla f(x_t) + \frac{1}{\eta}(x - x_t)\right]\Big|_{x = x_{t+1}} = 0$$

$$\Rightarrow x_{t+1} = x_t - \eta \nabla f(x_t)$$

Interpretation: We want to minimize the linear approx., without going too far w.r.t. $\frac{1}{2\eta}\|x - x_t\|_2^2$

More generally, we can plug in other sorts of distances

Warning: These will not necessarily be symmetric

def: Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a strictly convex and differentiable function. Then the associated Bregman divergence is

$$D_\phi(x, z) = \phi(x) - \phi(z) - \langle \nabla\phi(z), x - z \rangle$$

This is the difference between $\phi(x)$ and the estimate we get from the linear approx. at z

Think of it as locally quadratic

$$D_\phi(x, z) = (x - z)^T \nabla^2 \phi(y) (x - z)$$

for some point y depends on x and z

You can make this precise via remainder formulas for the Taylor expansion

We can now define MD more generally:

Mirror Descent

```
For $t = 1$ to $T$
    set $x_{t+1} =$
$\underset{x}{\operatorname{argmin}} \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{\eta} D_\phi(x, x_t) \right\}$
```

Some Important Examples

$\phi(x)$	$D_\phi(x, z)$
$\frac{\|x\|_2^2}{2}$	$\frac{\|x - z\|_2^2}{2}$
$\sum x_i \log x_i - \sum x_i$	$\sum x_i \log \frac{x_i}{z_i} - \sum x_i + \sum z_i$
"entropy"	"KL divergence"

$\vdots$ Many more in e.g. Dhillon, Tropp

Lemma 2 The iterates in MD satisfy

$$\nabla \phi(x_{t+1}) - \nabla \phi(x_t) = -\eta \nabla f(x_t)$$

Proof: As before, we can take the gradient of the expression we want to minimize and set it equal to zero at $x = x_{t+1}$:

$$\Rightarrow \nabla f(x_t) + \frac{1}{\eta} \nabla D_\phi(x, x_t)\Big|_{x=x_{t+1}} = 0$$

Using the definition of the Bregman divergence, we have:

$$\nabla f(x_t) + \frac{1}{\eta}\left(\nabla \phi(x) - \nabla \phi(x_t)\right)\Big|_{x=x_{t+1}} = 0$$

$$\Rightarrow \nabla \phi(x_{t+1}) - \nabla \phi(x_t) = -\eta \nabla f(x_t)$$

$\blacksquare$

The important thing is we can now characterize what point we reach, just as we did in the special case of Euclidean distance / vanilla GD

Consider $f(x) = \frac{\|Ax-b\|_2^2}{2}$

<u>Key Lemma</u>: When initialized at x_1, provided the step sizes are chosen so that MD converges to the minimum of f, we have:

$$x_t \longrightarrow \underset{x \text{ s.t. } Ax=b}{\operatorname{argmin}} \; D_\phi(x, x_1)$$

Bregman proj.

We will use the following standard fact:

Fact: $D_\phi(x, z)$ is convex in x

Proof of Key Lemma: First we write down the KKT conditions for the Bregman projection:

(1) x^* is the Bregman projection of x_1 onto the zero error set

$$\Updownarrow$$

(2) $\nabla D_\phi(x, x_1)\Big|_{x=x^*} \in \text{span}(A^T)$

i.e. if the direction of the linear approx. lives entirely in the space of constraints

Note: This implicitly uses the assumption that ϕ is differentiable

Now reusing our earlier calculation in the proof of Lemma 2, we have

$$\nabla D_\phi(x, x_1)\Big|_{x=x^*} = \nabla\phi(x^*) - \nabla\phi(x_1)$$

and applying Lemma 2 we have

$$\begin{aligned}
\nabla\phi(x^*) - \nabla\phi(x_1) &= \sum_{t=1}^{\infty} \nabla\phi(x_{t+1}) - \nabla\phi(x_t) \\
&= \sum_{t=1}^{\infty} -\eta \nabla f(x_t) \\
&= \sum_{t=1}^{\infty} -\eta A^T(Ax_t - b) \\
&\in \operatorname{span}(A^T)
\end{aligned}$$

∎

Note: To converge to a minimum of f, we sometimes use variable step sizes η_t that go to zero as t goes to infinity The argument goes thru with this change

Back to Reparameterization

Now let's return to our reparam. and figure out what kind of MD it is

Recall D_ϕ behaves like a locally quadratic function

$$D_\phi(x,z) = (x-z)^T \nabla^2\phi(y)(x-z)$$

The same way that we checked

$$D_\phi(x,z) = \frac{\|x-z\|_2^2}{2}$$

recovers vanilla GD, we can check that if we used

$$D_\phi(x, z) = \frac{(x-z)^T K (x-z)}{2}$$

for some p.d. matrix K we'd get

$$\left[\nabla f(x_t) + \frac{K(x - x_t)}{\eta}\right]\Bigg|_{x = x_{t+1}} = 0$$

$$\Rightarrow x_{t+1} = x_t - \eta K^{-1} \nabla f(x_t)$$

Thus for MD in general we'd get the update rule

$$x_{t+1} = x_t - \eta (\nabla^2 \phi(y))^{-1} \nabla f(x_t)$$

And as we take the step size to zero, $y \to x_t$ and we'd have:

$$\dot{x}(t) = -(\nabla^2 \phi(x(t)))^{-1} \nabla f(x(t))$$

And when $f(x) = \frac{\|Ax - b\|_2^2}{2}$ we'd get:

$$\dot{x}(t) = -(\nabla^2 \phi(x(t)))^{-1} A^T (Ax(t) - b)$$

This now looks familiar – we want some ϕ s.t.

$$(\nabla^2 \phi(x(t)))^{-1} = 4 D_x$$

where recall $D_x = \operatorname{diag}(x(t))$

<u>claim:</u> $\frac{d^2}{dx^2}(x \ln x - x) = \frac{1}{x}$

Thus we get that the following are equivalent for our choice of $f(x)$:

① gradient flow under the sq-param.

$$\Updownarrow$$

② MD under $\phi(x) \overset{\Delta}{=} \sum_i x_i \ln x_i - \sum_i x_i$

Note: MD with stepsize going to zero never leaves the nonnegative orthant because the divergence blows up

This is what we've been after all along:

"reparameterizations correspond to a change in local geometry"

Now let's complete the story by figuring out an implicit regularization problem that characterizes what we converge to:

Lemma 3: If $x_1 = \alpha \vec{1}$ then as $\alpha \to 0$ we have

$$x^* \longrightarrow \underset{x \text{ s.t. } \|Ax=b\|_1,\ x \in \mathbb{R}^d_{\geq 0}}{\operatorname{argmin}} \|x\|_1$$

Proof (sketch) Recall that for

$$\phi = \text{shannon entropy}$$

we get D_ϕ = KL divergence, i.e.

$$D_\phi(x,y) = \sum x_i \ln \frac{x_i}{y_i} - \sum x_i + \sum y_i$$

Now choosing $y = \alpha \vec{1}$ we get

$$D_\phi(x, \alpha\vec{1}) = \sum x_i \ln \frac{x_i}{\alpha} - \sum x_i + \alpha d$$

$$= \sum x_i \ln x_i + \left(\ln \frac{1}{\alpha} - 1\right) \|x\|_1 + \alpha d$$

And as $\alpha \to 0$ we have

$$D_\phi(x, \alpha\vec{1}) \to C_\alpha \|x\|_1$$

We need to be careful with the order of limits, but this can be done by showing that the limit of the gradient flow for any fixed α converges to an approximate minimizer of $\|\cdot\|_1$ where the gap goes to zero as α goes to zero ∎

Historical Notes: This connection btwn implicit regularization and MD was first discovered by Gunasekar, Lee, Soudry and Srebro

Also Gunasekar, Bhojanapalli, Neyshabur and Srebro made a bold Conjecture:

Conjecture [informal] In the noncommutative case $L_i : \mathbb{R}^{d \times d} \to \mathbb{R}$ gradient flow on

$$\sum_i \left(L_i(YY^T) - b_i \right)^2$$

converges to nuclear norm minimizer

$$X^* = \underset{\substack{X \text{ s.t. } L_i(X) = b_i \, \forall i \\ X \succeq 0}}{\operatorname{argmin}} \|X\|_* \leftarrow \text{sum of singular values}$$

What we just proved corresponds to the special case of diagonal matrices

Li, Ma and Zhang proved case where Li's satisfy matrix RIP

Theorem [Li, Luo, Lyu]: The conjecture is false (finds lower rank soln instead)

Further Results

Soudry, Hoffer, Nacson, Gunasekar and Srebro considered implicit regularization for SVMs:

Consider a dataset $D = \{(x_i, \underset{\pm 1}{\underset{\uparrow}{y_i}})\}$

def: we say D is linearly separable if $\exists\, a, b$ with

$$\operatorname{sgn}(a^T x_i + b) = y_i \quad \forall i$$

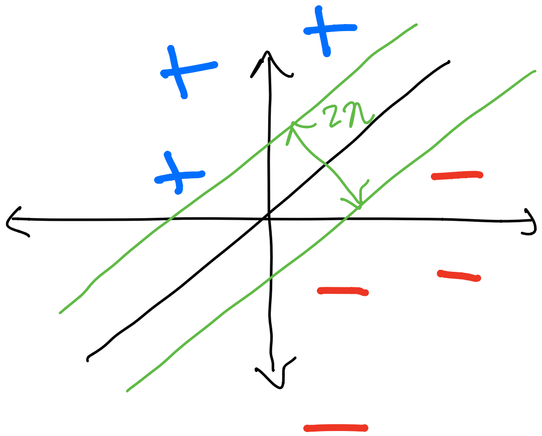
Note Can reduce to the case $b = 0$

When there is more than one sep. which one is the "best"?

def: We say a separator a, b has margin η if

$$\operatorname{sgn}(a^T x_i + b - \eta \|a\| y_i) = y_i$$

Pictorially, when a is a unit vector:



i.e. not only do we get all the examples right, but no example is close to the decision boundary

Main Question: What does gradient descent on a natural convex obj converge to?

Does it find the maximum margin linear separator?

Consider the usual gradient flow:

$$\dot{a}(t) = -\nabla \ell(a(t))$$

with the loss:

$$\ell(a) = \sum_i e^{-y_i a^T x_i}$$

Theorem [Soudry et al]: Let a^* be the unit vector in the direction of the maximum margin separator. Then

$$\frac{a(t)}{\|a(t)\|} \longrightarrow a^*$$

In particular $a(t) \sim a^* \log t + O(\log\log t)$

Taking a step back:

Key Question: Why is the maximum margin separator the "best"?

In a seminal work, Schapire, Freund, Bartlett and Lee proved a generalization bound for SVMs in terms of the margin

But does that mean it actually generalizes the best?

Or does it just optimizing the bounds we have?

Also Related

Theorem [Kawaguchi]: There are no spurrious local minima for

$$\sum_i \|A_1 A_2 \dots A_k x_i - b_i\|_2^2$$

linear deep net

Only bad saddle points for $k \geq 3$

Here bad saddle point means the Hessian has no negative eigenvalues

More Administration

New: I will have office hours Tues 3-4, zoom link is on Slack

Come in if you want to talk about anything theory of deep learning-related

New: HW #1 is now up on the course website

Grading for the course will be very lenient, but also good to practice the key material (often hard to find good hw problems in the theory of DL)