Quiz

1. Which of the following is NOT a name for the update rule $\boldsymbol{x}^{t+1} = \mathcal{P}(\boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t))$, in which $\mathcal{P}$ is a certain projection operator?

   A. regularized gradient descent B. first-order method
   C. projected gradient descent D. implicit regularization

2. Which "one" is left out in the leave-one-out trick for the general model $y_j = \psi_j^* \boldsymbol{H}^\natural \boldsymbol{X}^{\natural *} \phi_j (1 \leq j \leq m)$, in which each j corresponds to a collected sample?

   A. one dimension in each of the m equalities
   B. one of the above m samples
   C. one iteration in the course of gradient descent
   D. one of the eigenvectors in spectral initialization

Implicit Regularization in Nonconvex Statistical Estimation

Authors: Cong Ma, Kaizheng Wang, Yuejie Chi, Yuxin Chen

Presenter: Chengrun Yang (cy438)
Electrical and Computer Engineering, Cornell University

February 19, 2018

Nonlinear Systems: Problem Formulation

- want to know: $\boldsymbol{x}^\natural$
- data points collected: $\{y_j\}_{1\leq j\leq m}$
- relationship between measurements and ground truth: $y_j \approx \mathcal{A}_j(\boldsymbol{x}^\natural)$ $(1 \leq j \leq m)$, $\{A_j\}$ nonlinear map

Nonlinear Systems: Examples

Relationships among m measurements, design vectors and objects of interest $(1 \leq j \leq m)$:

1. phase retrieval: $y_j = (\boldsymbol{a}_j^\top \boldsymbol{x}^\natural)^2$
2. (noiseless) low-rank matrix completion:
   $Y_{j,k} = M^\natural{}_{j,k} = (\boldsymbol{X}^\natural \boldsymbol{X}^{\natural\top})_{j,k}$
3. blind deconvolution: $y_j = \boldsymbol{b}_j^* \boldsymbol{h}^\natural \boldsymbol{x}^{\natural *} \boldsymbol{a}_j$

A general model: $y_j = \psi_j^* \boldsymbol{H}^\natural \boldsymbol{X}^{\natural *} \phi_j$

What we often want...

- simple optimization algorithms, e.g., gradient descent
- known hyperparameters for these algorithms, e.g., no need to tune step size
- fast convergence, e.g., linear: $\mathrm{dist}(x^{t+1}, x^\natural) \leq (1-c)\mathrm{dist}(x^t, x^\natural)$, $c > 0$

Can we have them all?

Types of Gradient Descent

- vanilla: $x^{t+1} = x^t - \eta_t \nabla f(x^t)$
- regularized (NOT what we are going to consider here):
  - trimming/truncation: $x^{t+1} = x^t - \eta_t \mathcal{T}(\nabla f(x^t))$
  - regularized loss: $x^{t+1} = x^t - \eta_t(\nabla f(x^t) + \nabla R(x^t))$
  - projection: $x^{t+1} = \mathcal{P}(x^t - \eta_t \nabla f(x^t))$

What we care about: how to regularize in an **implicit** way.

Roadmap

linear convergence of gradient descent

$\Uparrow$

(local) strong convexity and smoothness

$\Uparrow$

region of incoherence and contraction

$\Uparrow$

leave-one-out trick

Background: Gradient Descent Theory

Definition: smoothness and strong convexity (SSC)

A twice continuously differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is:

- β-**smooth** if $\nabla^2 f(\boldsymbol{x}) \preceq \beta \boldsymbol{I}_n, \quad \exists \beta > 0, \forall \boldsymbol{x} \in \mathbb{R}^n$
- α-**strongly convex** if $\nabla^2 f(\boldsymbol{x}) \succeq \alpha \boldsymbol{I}_n, \quad \exists \alpha > 0, \forall \boldsymbol{x} \in \mathbb{R}^n$

A β-smooth and α-strongly convex (SSC) function f has condition number $\kappa := \beta / \alpha$.

Definition: relative ϵ-accuracy

The t-th iteration achieves relative ϵ-accuracy if $\left\|\boldsymbol{x}^t - \boldsymbol{x}^\natural\right\|_2 \leq \epsilon \left\|\boldsymbol{x}^0 - \boldsymbol{x}^\natural\right\|_2$.

Background: Gradient Descent Theory (continued)

Theorem: linear convergence under SSC

With step size $\eta_t \leq \frac{2}{\alpha+\beta}$, gradient descent on an SSC function has linear convergence $\left\| \boldsymbol{x}^{t+1} - \boldsymbol{x}^{\natural} \right\|_2 \leq c \left\| \boldsymbol{x}^{t} - \boldsymbol{x}^{\natural} \right\|_2$, in which $c = \sqrt{1 - \eta_t (\frac{2\alpha\beta}{\alpha+\beta})}$.

Specifically, when step size $\eta_t = \frac{2}{\alpha+\beta}$, gradient descent on an SSC objective function f converges to ϵ-accuracy within $O\left(\kappa \log \frac{1}{\epsilon}\right)$ iterations.

Phase Retrieval: Region of Incoherence and Contraction (RIC)

- $\left\| \boldsymbol{x} - \boldsymbol{x}^{\natural} \right\|_2 \leq \delta \left\| \boldsymbol{x}^{\natural} \right\|_2$ (neighborhood around optimum)
- $\max_{1 \leq j \leq m} \left| \boldsymbol{a}_j^{\top} \left(\boldsymbol{x} - \boldsymbol{x}^{\natural} \right) \right| \lesssim \sqrt{\log n} \left\| \boldsymbol{x}^{\natural} \right\|_2$ (incoherence)

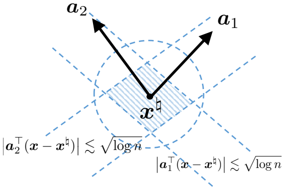


illustration of RIC

RIC Ensures Strong Convexity and Smoothness

With $\nabla^2 f(\boldsymbol{x}) = \frac{1}{m}\sum_{j=1}^{m}\left[3\left(\boldsymbol{a}_j^\top \boldsymbol{x}\right)^2 - y_j\right]\boldsymbol{a}_j\boldsymbol{a}_j^\top$,

Interpretation of Lemma 1 (Page 23; proof on Page 46-47)

With measurements $y_j = \left(\boldsymbol{a}_j^\top \boldsymbol{x}^\natural\right)^2$, $\boldsymbol{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$, $x^\natural \in \mathbb{R}^n$, $1 \le j \le m$, when

- $m \ge c_0 n \log n$
  $\rightarrow$ proximity between Hessian and its expectation
- $\left\|\boldsymbol{x} - \boldsymbol{x}^\natural\right\|_2 \le 2C_1$
  $\rightarrow$ further ensures positive definiteness of Hessian

the loss function is strongly convex; additionally, with

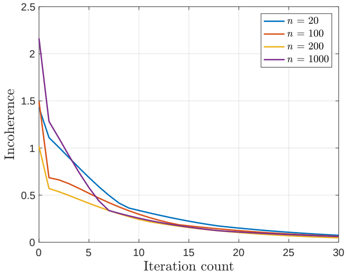
- $\max_{1\le j\le m}\left|\boldsymbol{a}_j^\top\left(\boldsymbol{x} - \boldsymbol{x}^\natural\right)\right| \le C_2\sqrt{\log n}$
  $\rightarrow$ upper bounds Hessian

the loss function is smooth.

Phase Retrieval: Implicit Regularization

Implicit regularization: Iterates automatically remain incoherent without explicit enforcement.

$$\frac{\max_{1\le j\le m}\left|\boldsymbol{a}_j^\top(\boldsymbol{x}^t-\boldsymbol{x}^\natural)\right|}{\sqrt{\log n}\left\|\boldsymbol{x}^\natural\right\|_2}$$



Cases of iterates (a) falling out of, or (b) staying within RIC

Now that RIC is good ...
How to start from and stay within RIC?

It turns out that ...

1. Spectral initialization guarantees in-RIC initialization
2. Leave-one-out trick ensures further iterates remain in RIC

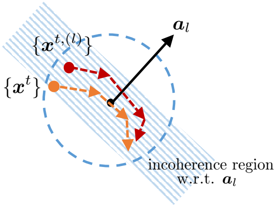
Phase Retrieval: The Leave-One-Out Trick

In phase retrieval, $\{\boldsymbol{x}^{t,(l)}\}$ is a sequence of auxiliary iterates for

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x}) = \frac{1}{m}\sum_{j=1, j\neq l}^{m}\left[\left(\boldsymbol{a}_j^\top \boldsymbol{x}\right)^2 - y_j\right]^2$$

For each $1 \leq l \leq m$, $\{\boldsymbol{x}^{t,(l)}\}$ satisfies:

- $\boldsymbol{x}^t \approx \boldsymbol{x}^{t,(l)}, \qquad t \geq 0$
- $\left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^\natural\right)\right| \lesssim \sqrt{\log n}\,\left\|\boldsymbol{x}^\natural\right\|_2$
  easy to satisfy



Thus we can show the original iterates $\{\boldsymbol{x}^t\}$ fall in RIC.

Note: This leave-one-out optimization is never performed!

Phase Retrieval: Spectral Initialization

Spectral Initialization

Given m quadratic equations $y_j = \left(\boldsymbol{a}_j^\top \boldsymbol{x}^\natural\right)^2$ $(j = 1, \ldots, m)$, set $\boldsymbol{x}^0 = \sqrt{\lambda_1(\boldsymbol{Y})/3}\ \tilde{\boldsymbol{x}}^0$, in which $\lambda_1(\boldsymbol{Y})$ and $\tilde{\boldsymbol{x}}^0$ are the leading eigenvalue and eigenvector of $\boldsymbol{Y} = \frac{1}{m}\sum_{j=1}^m y_j \boldsymbol{a}_j \boldsymbol{a}_j^\top$.

Idea:

- leave-one-out iterates $\boldsymbol{x}^{t,(l)}$ are incoherent
- spectral initialization $\boldsymbol{x}^0$ is not far from $\boldsymbol{x}^{0,(l)}$ (Davis-Kahan), and is thus also incoherent

The General Setting

- collected samples: $y_j = \psi_j^* \boldsymbol{H}^\natural \boldsymbol{X}^{\natural *} \phi_j \ (1 \le j \le m)$
- empirical loss:
  $f(\boldsymbol{Z}) := f(\boldsymbol{H}, \boldsymbol{X}) = \frac{1}{m} \sum_{j=1}^{m} \left| \psi_j^* \boldsymbol{H} \boldsymbol{X}^* \phi_j - y_j \right|^2$
- incoherence: $\max_j \left\| \phi_j^* (\boldsymbol{X} - \boldsymbol{X}^\natural) \right\|_2$ and $\max_j \left\| \psi_j^* (\boldsymbol{H} - \boldsymbol{H}^\natural) \right\|_2$ are upper bounded, which leads to RIC
- SSC: $\boldsymbol{0} \prec \alpha \boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{Z}) \preceq \beta \boldsymbol{I}, \qquad \forall \boldsymbol{Z} \in \text{RIC}$

Leave-One-Out Establishes Incoherence

- Goal: upper bound $\left\|\phi_l^*\left(\boldsymbol{X}^{t+1}-\boldsymbol{X}^{\natural}\right)\right\|_2$ to ensure incoherence
- Approach: construct auxiliary sequence $\{\boldsymbol{Z}^{t,(l)}\}=\{(\boldsymbol{X}^{t,(l)},\boldsymbol{H}^{t,(l)})\}$ such that $\boldsymbol{X}^{t,(l)}$ (resp. $\boldsymbol{H}^{t,(l)}$) is independent of any sample involving ϕ_l (resp. ψ_l)

$$\Rightarrow \left\|\phi_l^*\left(\boldsymbol{X}^{t+1}-\boldsymbol{X}^{\natural}\right)\right\|_2 \leq \left\|\phi_l\right\|_2 \underbrace{\left\|\boldsymbol{X}^{t+1}-\boldsymbol{X}^{t+1,(l)}\right\|_{\mathrm{F}}}_{①} + \underbrace{\left\|\phi_l^*\left(\boldsymbol{X}^{t+1,(l)}-\boldsymbol{X}^{\natural}\right)\right\|_2}_{②}$$

①: proximity between original and leave-one-out iterates
②: incoherence of leave-one-out iterates

Strong Convexity and Smoothness of Statistical Estimation Problems

The strong convexity and smoothness of Hessian holds with high probability in the following regions:

- phase retrieval: near global optimal $X^\natural$, under the incoherence condition
- low-rank matrix completion and blind deconvolution: near global optimal $X^\natural$, under the incoherence condition, along certain directions

Open Questions

The vanilla gradient descent achieves implicit regularization for some statistical estimation problems.

- Is incoherence always guarenteed (e.g., in the case of non-Gaussian design)?
- What other problems might this framework be applied to? e.g., other statistical estimation problems, or further, constrained optimization problems.