CS-GY 9223 I: Lecture 8
Coordinate decent and non-convex models.

NYU Tandon School of Engineering, Prof. Christopher Musco

Main idea: Trade slower convergence (more iterations) for cheaper iterations.

Stochastic Coordinate Descent: Only compute a single random entry of $\nabla f(\mathbf{x})$ on each iteration:

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \frac{\partial f}{\partial x_2}(\mathbf{x}) \\ \vdots \\ \frac{\partial f}{\partial x_d}(\mathbf{x}) \end{bmatrix} \qquad \nabla_i f(\mathbf{x}) = \begin{bmatrix} 0 \\ \frac{\partial f}{\partial x_i}(\mathbf{x}) \\ \vdots \\ 0 \end{bmatrix}$$

Update: $\underline{\mathbf{x}^{(t+1)}} \leftarrow \underline{\mathbf{x}^{(t)}} + \eta \nabla_i f(\mathbf{x}^{(t)})$.

When $\mathbf{x}$ has d parameters, computing $\nabla_i f(\mathbf{x})$ often costs just a $1/d$ fraction of what it costs to compute $\nabla f(\mathbf{x})$

Example: $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ for $\mathbf{A} \in \mathbb{R}^{n \times d}, \mathbf{x} \in \mathbb{R}^d, \mathbf{b} \in \mathbb{R}^n$.

- $\nabla f(\mathbf{x}) = 2\mathbf{A}^T\mathbf{A}\mathbf{x} - 2\mathbf{A}^T\mathbf{b}$.
- $\nabla_i f(\mathbf{x}) = 2\left[\mathbf{A}^T\mathbf{A}\mathbf{x}\right]_i - 2\left[\mathbf{A}^T\mathbf{b}\right]_i$.

Full gradient for:
$f(x) = \|Ax - b\|_2^2$
$\to O(nd)$ time

- $\mathbf{A}\mathbf{x}^{(t+1)} = \mathbf{A}\left(\mathbf{x}^{(t)} + c \cdot \mathbf{e}_i\right)$ — $O(n)$ time
- $2\left[\mathbf{A}^T\left(\mathbf{A}\mathbf{x}^{(t+1)} - \mathbf{b}\right)\right]_i$ — $O(n)$ time

STOCHASTIC COORDINATE DESCENT

Stochastic Coordinate Descent:

- Choose number of steps T and step size η.
- For $t = 1, \ldots, T$:
    - Pick random $j \in 1, \ldots, d$.
    - $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta \nabla_j f(\mathbf{x}^{(i)})$
- Return $\hat{\mathbf{x}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}^{(t)}$.

Theorem (Stochastic Coordinate Descent convergence)

Given a G-Lipschitz function f with minimizer $\mathbf{x}^*$ *and initial point* $\mathbf{x}^{(1)}$ *with* $\|\mathbf{x}^{(1)} - \mathbf{x}^*\|_2 \leq R$*, SCD with step size* $\eta = \frac{1}{Rd}$ *satisfies the guarantee:*

$$\mathbb{E}[f(\hat{\mathbf{x}}) - f(\mathbf{x}^*)] \leq \frac{2GR}{\sqrt{T/d}}$$

$$\leq O\left(\frac{GR}{\sqrt{T}}\right)$$ for full gradient descent

How can we improve on this?

SCD takes $d\times$ more iterations for same error.

$\|Ax-b\|_2^2$ $\qquad$ $Ax \approx b$

Often it doesn't make sense to sample i uniformly at random:

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & -.5 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 & 0 \end{bmatrix} \quad \underbrace{\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}}_{x} \qquad \mathbf{b} = \begin{bmatrix} 10 \\ 42 \\ -11 \\ -51 \\ 34 \\ -22 \end{bmatrix}$$

Select indices i proportional to $\|\mathbf{a}_i\|_2^2$:

ith column of A

$$\Pr[\text{select index } i \text{ to update}] = \frac{\|\mathbf{a}_i\|_2^2}{\underbrace{\sum_{j=1}^{d} \|\mathbf{a}_j\|_2^2}_{= \|A\|_F^2}} = \frac{\|\mathbf{a}_i\|_2^2}{\|\mathbf{A}\|_F^2}$$

Let's analyze this approach.

Specialization of SCD to $\|Ax - b\|_2^2$:

Random Kaczmarz method

Randomized Coordinate Descent (Strohmer, Vershynin 2007 / Leventhal, Lewis 2018)

- For iterate $x^{(t)}$, let $r^{(t)}$ be the residual:

$f(x^{(t)}) = \|r^{(t)}\|_2^2$

$$r^{(t)} = Ax^{(t)} - b$$

- $x^{(t+1)} = x^{(t)} - ce_j$. Here c is a scalar and e_j is a standard basis vector.

e_j

- $r^{(t+1)} = r^{(t)} - ca_j$. Here a_j is the i^{th} column of A.

$$Ax^{(t+1)} - b = A(x^{(t)} - ce_j) - b$$
$$= r^{(t)} - ca_j$$

Typically: η = fixed. "line search" $\rightarrow f(x^{(t+1)})$

What choice for c minimizes $\|\mathbf{r}^{(t+1)}\|_2^2$?

$x^{(t+1)} = x^{(t)} - c e_j$ so that I minimize $f(x^{(t+1)})$

$\|r^{(t+1)}\|_2^2 = \|r^{(t)} - c a_j\|_2^2$

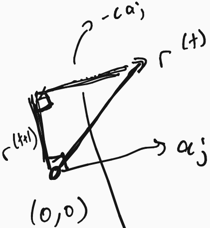
$g(c) = \|r^{(t)} - c a_j\|_2^2$

$= \|r^{(t)}\|_2^2 - 2c a_j^T r^{(t)} + c^2 \|a_j\|_2^2$

$g'(c) = -2 a_j^T r^{(t)} + 2c \|a_j\|_2^2$

$= 0$ when

$$\boxed{c = \frac{a_j^T r^{(t)}}{\|a_j\|_2^2}}$$



$\rightarrow$ "project $r^{(t)}$ onto a_j."

$\frac{a_j^T r^{(t)}}{\|a_j\|_2^2} \cdot a_j$

Specialization of SCD to $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$:

$x^{(1)} \ldots\ldots x^{(t)}$

Randomized Coordinate Descent

- Choose number of steps T.
- Let $\mathbf{x}^{(1)} = \mathbf{0}$ and $\mathbf{r}^{(1)} = \mathbf{b}$.
- For $t = 1, \ldots, T$:
    - Pick random $j \in 1, \ldots, d$. Index j is selected with probability proportional to $\|\mathbf{a}_j\|_2^2 / \|A\|_F^2$.
    - Set $c = \mathbf{a}_j^T \mathbf{r}^{(t)} / \|\mathbf{a}_j\|_2^2$
    - $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - c\mathbf{e}_j$
    - $\mathbf{r}^{(t+1)} = \mathbf{r}^{(t)} - c\mathbf{a}_j$
- Return $\mathbf{x}^{(T)}$.

$\|A\|_F^2 = \sum_{ij} (A_{ij})^2$

Frobenius

Claim

$$\|r^{(t+1)}\|_2^2 + \|c\,a_i\|_2^2 = \|r^{(t)}\|_2^2$$

$$\mathbb{E}\|\mathbf{r}^{(t+1)}\|_2^2 = \|\mathbf{r}^{(t)}\|_2^2 - \frac{1}{\|\mathbf{A}\|_F^2}\|\mathbf{A}^T\mathbf{r}^{(t)}\|_2^2$$

$$\|r^{(t+1)}\|_2^2 = \|r^{(t)} - c\,a_i\|_2^2 = \|r^{(t)}\|_2^2 - c^2\|a_i\|_2^2$$

$$= \|r^{(t)}\|_2^2 - \frac{(a_i^T r^{(t)})^2}{\|a_i\|_2^4}\|a_i\|_2^2$$

$$= \|r^{(t)}\|_2^2 - \frac{(a_i^T r^{(t)})^2}{\|a_i\|_2^2}$$

$$\mathbb{E}[\|r^{(t+1)}\|_2^2] = \sum_{i=1}^{d} \frac{\cancel{\|a_i\|_2^2}}{\|A\|_F^2}\left(\|r^{(t)}\|_2^2 - \frac{(a_i^T r^{(t)})^2}{\cancel{\|a_i\|_2^2}}\right)$$

$$= \|r^{(t)}\|_2^2 - \frac{1}{\|A\|_F^2}\underbrace{\sum_{i=1}^{d}(a_i^T r^{(t)})^2}_{\|A^T r^{(t)}\|_2^2}$$

Any residual $\mathbf{r}$ can be written as $\mathbf{r} = \mathbf{r}^* + \bar{\mathbf{r}}$ where $\mathbf{r}^* = \mathbf{A}\mathbf{x}^* - \mathbf{b}$ and $\bar{\mathbf{r}} = \mathbf{A}(\mathbf{x}^t - \mathbf{x}^*)$. Note that $\underbrace{\mathbf{A}^T\mathbf{r}^* = 0}$ and $\bar{\mathbf{r}} \perp \mathbf{r}^*$. $\underbrace{(x^{(t)} - x^*)^T A^T}_{\bar{r}^T} r^* = 0$

Claim $A^T A x^* - A^T b = 0 \quad \nabla f(x^*) = 0$

$$\mathbb{E}\|\bar{\mathbf{r}}^{(t+1)}\|_2^2 = \|\bar{\mathbf{r}}^{(t)}\|_2^2 - \frac{1}{\|\mathbf{A}\|_F^2}\|\mathbf{A}^T\bar{\mathbf{r}}^{(t)}\|_2^2$$

$$\leq \|\bar{\mathbf{r}}^{(t)}\|_2^2 - \frac{\lambda_{\min}(\mathbf{A}^T\mathbf{A})}{\|\mathbf{A}\|_F^2}\|\bar{\mathbf{r}}^{(t)}\|_2^2$$

$$= \left(1 - \frac{\lambda_{\min}(A^T A)}{\|A\|_F^2}\right)\|\bar{r}^{(t)}\|_2^2$$

$\mathbb{E}\|r^{(t+1)}\|_2^2 = \|r^*\|_2^2 + \mathbb{E}[\|\bar{r}^{(t+1)}\|_2^2$ $\quad \|A^T\bar{r}^{(t)}\|_2^2 \geq \lambda_{\min}(A^T A)\|\bar{r}^{(t)}\|_2^2$

$\|\bar{r}^{(t)}\|_2^2 = \|r^*\|_2^2 + \|\bar{r}^{(t)}\|_2^2$

$\|A^T r^{(t)}\|_2^2 = \|A^T\bar{r}^{(t)} + \underbrace{A^T r^*}_{\to 0}\|_2^2 = \|A^T\bar{r}^{(t)}\|_2^2$

Theorem (Randomized Coordinate Descent convergence)

After T steps of RCD with importance sampling run on $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$, we have:

$\to \|r^{(t)}\|_2^2 - \|r^*\|_2^2 = \|\bar{r}^{(t)}\|_2^2$

T

$$\mathbb{E}[f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*)] \leq \left(1 - \frac{\lambda_{\min}(\mathbf{A}^T\mathbf{A})}{\|\mathbf{A}\|_F^2}\right)^T [f(\mathbf{x}^{(0)}) - f(\mathbf{x}^*)]$$

Corollary: After $T = O(\frac{\|\mathbf{A}\|_F^2}{\lambda_{\min}(\mathbf{A}^T\mathbf{A})} \log \frac{1}{\epsilon})$ we obtain error $\epsilon\|\mathbf{b}\|_2^2$.

Is this more or less iterations than the $\boxed{T = O(\frac{\lambda_{\max}(\mathbf{A}^T\mathbf{A})}{\lambda_{\min}(\mathbf{A}^T\mathbf{A})} \log \frac{1}{\epsilon})}$ required for gradient descent to converge?

$$\|\mathbf{A}\|_F^2 = \mathrm{tr}(\mathbf{A}^T\mathbf{A}) = \sum_{i=1}^{d} \lambda_i(\mathbf{A}^T\mathbf{A})$$

$$\lambda_{\max}(\mathbf{A}^T\mathbf{A}) \leq \|\mathbf{A}\|_F^2 \leq d \cdot \lambda_{\max}(\mathbf{A}^T\mathbf{A})$$

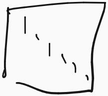
For solving $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$,

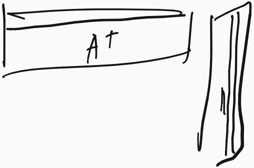
$$(\#\text{ GD Iterations}) \leq (\#\text{ RCD Iterations}) \leq d \cdot (\#\text{ GD Iterations})$$

But RCD iterations are cheaper by a factor of d.

When does $\|A\|_F^2 = \mathrm{tr}(A^T A) = d \cdot \lambda_{\max}(A^T A)$?

orthonormal columns

$A^T A =$ 

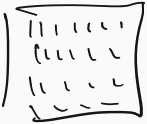


For all i, j

$a_j^T a_i = 0$

$a_j^T a_i = \text{small}$

When does $\|A\|_F^2 = \mathrm{tr}(A^T A) = 1 \cdot \lambda_{\max}(A^T A)$?



$= A^T A$

$\|A\|_F^2 = \lambda_{\max}(A^T A) = d^2$

$a_i^T a_j = 1$ for all i, j

similar vectors

all ones matrix

Roughly:

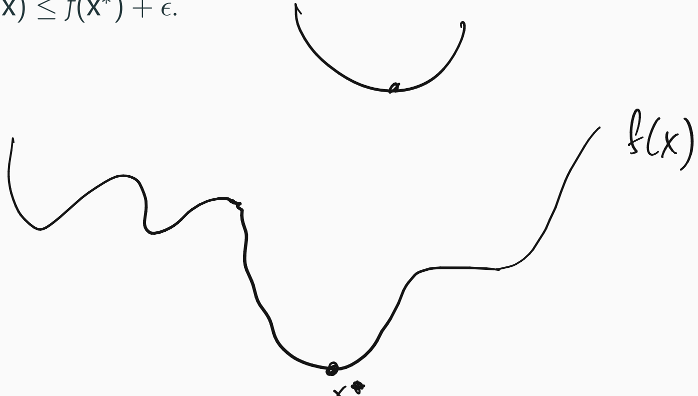
Stochastic Gradient Descent performs well when data points (rows) are repetitive.

Stochastic Coordinate Descent performs well when data features (columns) are repetitive.

NON-CONVEX OPTIMIZATION

VISUALIZATION

Given $f(\mathbf{x})$ which is potentially non-convex, find $\hat{\mathbf{x}}$ such that $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.



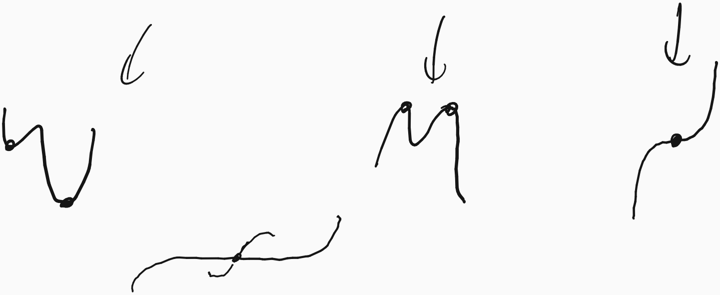
We understand very little about optimizing non-convex functions in comparison to convex functions, but not nothing. In many cases, we're still figuring out the right questions to ask.

Definition (Stationary point)

For a differentiable function f, a stationary point is any $\mathbf{x}$ with:

$$\nabla f(\mathbf{x}) = \mathbf{0} \qquad \nabla f(x^*) = 0$$

local/global minima - local/global maxima - saddle points



Reasonable goal: Find an approximate stationary point $\hat{\mathbf{x}}$ with

$$\|\nabla f(\hat{\mathbf{x}})\|_2 \leq \epsilon.$$

SMOOTHNESS FOR NON-CONVEX FUNTIONS

Definition

A differentiable (potentially non-convex) function f is β smooth if for all $\mathbf{x}, \mathbf{y}$,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq \beta \|\mathbf{x} - \mathbf{y}\|_2$$

Corollary: For all $\mathbf{x}, \mathbf{y}$

$$\left|\nabla f(\mathbf{x})^T(\mathbf{x} - \mathbf{y}) - [f(\mathbf{x}) - f(\mathbf{y})]\right| \leq \frac{\beta}{2}\|\mathbf{x} - \mathbf{y}\|_2^2.$$

Theorem

If Gradient Descent is run with step size $\eta = \frac{1}{\beta}$ on a differentiable function f with global minimum $\mathbf{x}^*$ then after $T = O(\frac{\beta[f(\mathbf{x}^{(1)}) - f(\mathbf{x}^*)]}{\epsilon})$ we will find an ϵ-approximate stationary point $\hat{\mathbf{x}}$.

1. $\nabla f(x^{(t)})^T \underbrace{(x^{(t)} - x^{(t+1)})}_{= \eta \nabla f(x^{(t)})} - f(x^{(t)}) + f(x^{(t+1)}) \leq \frac{\beta}{2} \|\underbrace{x^{(t)} - x^{(t+1)}}_{= \eta \nabla f(x^{(t)})}\|_2^2$

2. $f(x^{(t+1)}) - f(x^{(t)}) \leq \frac{\beta}{2} \eta^2 \|\nabla f(x^{(t)})\|_2^2 - \eta \|\nabla f(x^{(t)})\|_2^2$

$= -\frac{\eta}{2} \|\nabla f(x^{(t)})\|_2^2$ since $\beta = 1/\eta$

3. $\frac{1}{T} \sum_{t=1}^{T} \frac{\eta}{2} \|\nabla f(x^{(t)})\|_2^2 \leq \frac{1}{T} \sum_{t=1}^{T} f(x^{(t)}) - f(x^{(t+1)})$

4. $\min_t \|\nabla f(x^{(t)})\|_2^2 \cdot \frac{\eta}{2} \leq \frac{1}{T} [f(x^{(1)}) - \underbrace{f(x^{(T+1)})}_{\geq f(x^*)}]$ $\longrightarrow$

Theorem

If Gradient Descent is run with step size $\eta = \frac{1}{\beta}$ on a differentiable function f with global minimum $\mathbf{x}^*$ then after $T = O(\frac{\beta[f(\mathbf{x}^{(1)}) - f(\mathbf{x}^*)]}{\epsilon})$ we will find an ϵ-approximate stationary point $\hat{\mathbf{x}}$.

(cont.) Let $\hat{x} = \arg\min_t \|\nabla f(x^{(t)})\|_2$.

$$\|\nabla f(\hat{x})\|_2^2 \leq \frac{2}{T\eta}[f(x^{(1)}) - f(x^*)]$$

$$= \frac{2\beta}{T}.$$

Setting $T = \frac{2\beta}{\varepsilon} \cdot [f(x^{(1)}) - f(x^*)]$ gives the bound.

If GD can find a stationary point and that seems to work for your problem, are there algorithms which find a stationary point faster using preconditioning, acceleration, stocastic methods, etc.?

What if my function only has global minima and stationary points? Randomized methods (SGD, perturbed gradient methods, etc.) can "escape" stationary points under some minor assumptions.

Example: $\min_{\mathbf{x}} \frac{-\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$

function w/ only saddles + minima

- **Global minimum**: Top eigenvector of $\mathbf{A}^T\mathbf{A}$ (i.e., top principal component of $\mathbf{A}$).
- **Stationary points:** All other eigenvectors of $\mathbf{A}$.

Useful for lots of other matrix factorization problems beyond vanilla PCA.

QUESTIONS IN NON-CONVEX OPTIMIZATION

- Can random or careful initialization lead to a good minima?
- Can we escape “shallow” local minima.
- Is a global minima even needed?