Optimization for well-behaved problems

For statistical learning problems, "well-behaved" means:

- signal to noise ratio is decently high
- correlations between predictor variables are under control
- number of predictors p can be larger than number of observations n, but not absurdly so

For well-behaved learning problems, people have observed that gradient or generalized gradient descent can converge extremely quickly (much more so than predicted by $O(1/k)$ rate)

Largely unexplained by theory, topic of current research. E.g., very recent work[4] shows that for some well-behaved problems, w.h.p.:

$$\|x^{(k)} - x^\star\|^2 \leq c^k \|x^{(0)} - x^\star\|^2 + o(\|x^\star - x^{\text{true}}\|^2)$$

[4] Agarwal et al. (2012), *Fast global convergence of gradient methods for high-dimensional statistical recovery*

Administrivia

- HW2 out as of this past Tuesday—due 10/9
- Scribing
  - Scribes 1–6 ready soon; handling errata
  - missing days: 11/6, 12/4, and 12/6
- Projects:
  - you should expect to be contacted by TA mentor in next weeks
  - project milestone: 10/30

Matrix differential calculus

10-725 Optimization

Geoff Gordon
Ryan Tibshirani

Matrix calculus pain

- Take derivatives of fns involving matrices:
  - write as huge multiple summations w/ lots of terms
  - take derivative as usual, introducing more terms
    - case statements for $i = j$ vs. $i \neq j$
  - try to recognize that output is equivalent to some human-readable form
  - hope for no indexing errors…
- Is there a better way?

Differentials

- Assume f sufficiently "nice"
- Taylor: f(y) = f(x) + f'(x) (y–x) + r(y–x)
  - with r(y–x) / |y–x| → 0 as y → x x fixed
- Notation: $df = f(y) - f(x)$ $dx = y - x$ "differentials"

$df = f'(x)\,dx\ \cancel{[+ r(dx)]}$

Definition

- Write
  - $dx = y - x$
  - $df = f(y) - f(x)$
- Suppose
  - df = $a(x; dx) + r(dx)$
  - with a linear in dx
  - and $r(dx)/|dx| \to 0$ as $dx \to 0$
- Then: $a(x; dx)$ is differential of f

$a(x; k\,dx) = k\,a(x; dx)$

$a(x; dx_1 + dx_2) = a(x; dx_1) + a(x; dx_2)$

Matrix differentials

- For matrix X or matrix-valued function F(X):
  - dX = $[dX_{ij}]_{ij}$
  - dF = $a(X; dX) + r(dX)$
  - where a linear in dX
  - and $r(dX)/\|dX\| \to 0$ as $dX \to 0$
- Examples: $f: \mathbb{R}^n \to \mathbb{R}^m$ $\quad df = \underbrace{J(x)}_{\mathbb{R}^{n \times m}} dx$

$f: \mathbb{R}^n \to \mathbb{R}$ $\quad df = \underbrace{g(x)}_{\mathbb{R}^n}{}^T dx$ $\quad dg = \underbrace{H(x)}_{\mathbb{R}^{n \times n}} dx$

Working with differentials

- Linearity:
  - d(f(x) + g(x)) = $df(x) + dg(x)$
  - d(k f(x)) = $k\, df(x)$
- If g linear, dg(f(x)) = g(df(x)); for example
  - reshape(A, [m n k …])
  - vec(A) = A(:) = reshape(A, [], 1)
  - tr(A) = $\sum_i A_{ii}$
  - A^T

Reshape

```
>> A = reshape(1:24, [2 3 4])

A(:,:,1) =
     1     3     5
     2     4     6

A(:,:,2) =
     7     9    11
     8    10    12

A(:,:,3) =
    13    15    17
    14    16    18

A(:,:,4) =
    19    21    23
    20    22    24
```

```
>> B = reshape(A, [4 3 2])

B(:,:,1) =
     1     5     9
     2     6    10
     3     7    11
     4     8    12

B(:,:,2) =
    13    17    21
    14    18    22
    15    19    23
    16    20    24
```

Working with differentials

- Chain rule: L(x) = f(g(x))
  - want: $dL(x) = f'(g(x))g'(x)\,dx$
  - have: $df = a(g(x); dg) \; [+ r(dg)] \rightarrow f'(g(x))\,dg$
    $dg = b(x; dx) \; [+ s(dx)] \rightarrow g'(x)\,dx$

$r(dg)/\|dg\| \to 0$ as $dg \to 0$

$a(g(x); dg)$

$$dL = a(g(x); b(x; dx) + s(dx)) + r(dg)$$

$$= \underline{a(g(x); b(x; dx))} + [a(g(x); s(dx)) + r(dg)]$$

$$a(g(x); s(dx))/\|dx\| + r(dg)/\|dx\| \to 0 \text{ as } dx \to 0$$

$$a\left(g(x); \frac{s(dx)}{\|s(dx)\|}\right) \frac{\|s(dx)\|}{\|dx\|} + r(dg)/\|dg\| \; \frac{\|dg\|}{\|dx\|}$$

Working with differentials

- Product rule: L(x) = c(f(x), g(x))
  - where c is ***bilinear*** = linear in each argument (with other argument fixed)
  - e.g., L(x) = f(x)g(x): f, g scalars, vectors, or matrices

$$dL = c(df; g(x)) + c(f(x); dg)$$

Lots of products

- Cross product: $d(a \times b) =$ $da \times b + a \times db$
- Hadamard product $A \circ B = A .* B$ ← commutative
  - $(A \circ B)_{ij} =$ $A_{ij} B_{ij}$
  - $d(A \circ B) =$ $dA \circ B + A \circ dB$
- Kronecker product $d(A \otimes B) =$ $dA \otimes B + A \otimes dB$
- Frobenius product $A{:}B =$ $\sum_{ij} A_{ij} B_{ij}$
- Khatri-Rao product: $d(A*B) =$ $dA * B + A * dB$

Kronecker product

```
>> A = reshape(1:6, 2, 3)

A =

     1     3     5
     2     4     6

>> B = 2*ones(2)

B =

     2     2
     2     2
```

```
>> kron(A, B)

ans =

     2     2     6     6    10    10
     2     2     6     6    10    10
     4     4     8     8    12    12
     4     4     8     8    12    12

>> kron(B, A)

ans =

     2     6    10     2     6    10
     4     8    12     4     8    12
     2     6    10     2     6    10
     4     8    12     4     8    12
```

Hadamard product

- a, b vectors
  - $a \circ b =$ $\mathrm{diag}(a)\, b = \mathrm{diag}(b)\, a$
  - $\mathrm{diag}(a)\ \mathrm{diag}(b) =$ $\mathrm{diag}(a \circ b)$
  - $\mathrm{tr}(\mathrm{diag}(a)\ \mathrm{diag}(b)) =$ $a^T b$
  - $\mathrm{tr}(\mathrm{diag}(b)) =$ $\mathrm{tr}(\mathrm{diag}(1)\ \cancel{b}) = 1^T b$

    $\mathrm{diag}(b)$

Some examples

- $L = (Y-XW)^T(Y-XW)$: differential wrt W
  - $dL = d(Y^TY - Y^TXW - W^TX^TY + W^TX^TXW)$

    $= -Y^TX\,dW - dW^TX^TY + dW^TX^TXW + W^TX^TX\,dW$

Some examples

matrix X

- L = $[X_{ij}]^2/2$ dL = $\begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & & X_{ij} & 0 \\ 0 & \cdots & & 0 \end{bmatrix} \circ dX$
- L = $[X^T X]_{ij}^2/2$
  - dL = $\begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & & (X^T X)_{ij} & \vdots \\ 0 & \cdots & & 0 \end{bmatrix} \circ d(X^T X)$

$= M \circ (dX^T X + X^T dX)$

Trace

- tr(A) = $\sum_i A_{ii}$
  - d tr(f(x)) = $tr(d f(x))$
  - tr(x) = x ← scalar
  - tr(X^T) = $tr(X)$

$\sum_k A_{ki} B_{kj}$

$(A^T B)_{ij} = \cancel{\sum_k A_{ik} B_{jk}}$

$\sum_i (A^T B)_{ii} = \cancel{\sum_i \sum_k A_{ik} B_{ik}}$

$\sum_i \sum_k A_{ki} B_{ki}$

- Frobenius product:
  - A:B = $\sum_{ij} A_{ij} B_{ij} = tr(A^T B)$

Trace rotation

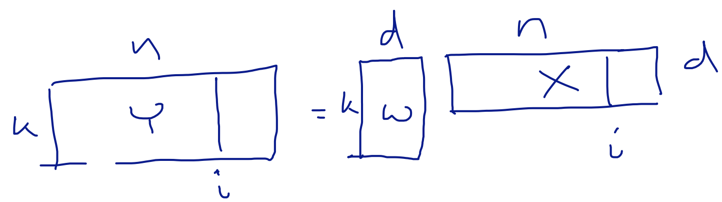
- tr(AB) = $\mathrm{tr}(BA)$
- tr(ABC) = $\mathrm{tr}(CAB) = \mathrm{tr}(BCA)$
  - size(A): $m \times n$
  - size(B): $n \times k$
  - size(C): $k \times m$

More

- Identities: for a matrix X,
  - $d(X^{-1}) =$ $X^{-1}(dX)X^{-1}$
  - $d(\det X) =$ $(\det X)\,\mathrm{tr}(X^{-1}dX)$
  - $d(\ln|\det X|) =$ $\mathrm{tr}(X^{-1}dX)$
  - …

Example: linear regression

- Training examples: $(x_i, y_i) \quad i = 1 \ldots n$
- Input feature vectors: $x_i \in \mathbb{R}^d$
- Target vectors: $y_i \in \mathbb{R}^k$



- Weight matrix: $W \in \mathbb{R}^{k \times d}$
- $\min_W L = \sum_i \|y_i - W x_i\|_2^2$
  - as matrix: $\|Y - WX\|_F^2$

I accidentally transposed WX->XW here, compared to the previous slide -- so for this slide only, the regression is from rows of X to rows of Y

can do it either way...

Linear regression

XW

- $L = ||Y - \cancel{WX}||_F^2 = (Y - XW) : (Y - XW)$

$= tr((Y - XW)^T (Y - XW))$

$= tr(Y^T Y - Y^T X W - W^T X^T Y + W^T X^T X W)$

- $dL = -tr(Y^T X\, dW + \underbrace{dW^T X^T Y}_{Y^T X\, dW}) + d\,tr(W^T X^T X W)$

$tr(dW^T X^T X W + W^T X^T X\, dW)$

$= -2 tr(Y^T X\, dW) + 2 tr(W^T X^T X\, dW)$

$Y^T X = W^T X^T X$

$Y^T X (X^T X)^{-1} = W^T$

Identification theorems

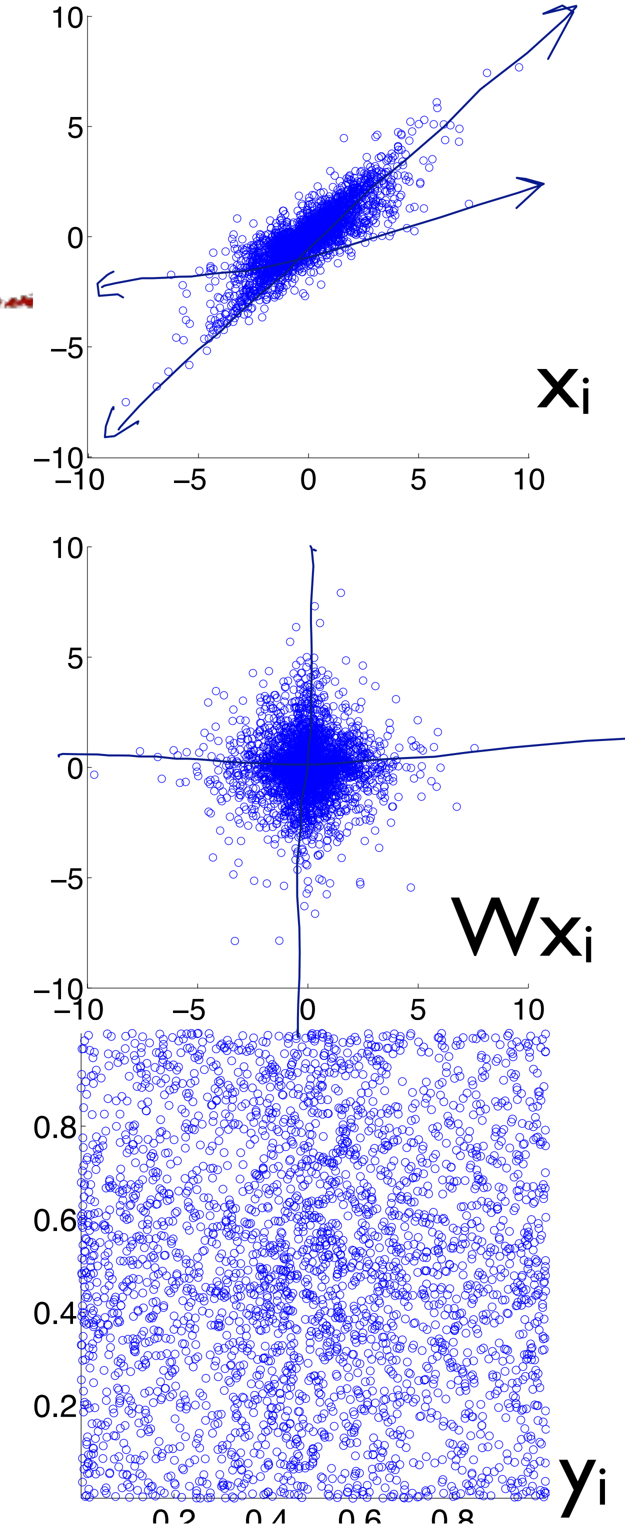
- Sometimes useful to go back and forth between differential & ordinary notations
  - not always possible: e.g., $d(X^T X) =$ $dX^T X + X^T dX$
- Six common cases (ID thms):

df/dx velocity gradient Jacobian

ID for df(x)	scalar x	vector ***x***	matrix X
scalar f	$df = a\,dx$	$df = \boldsymbol{a}^T d\boldsymbol{x}$	$df = \mathrm{tr}(A^T dX)$
vector ***f***	$d\boldsymbol{f} = \boldsymbol{a}\,dx$	$d\boldsymbol{f} = A\,d\boldsymbol{x}$	
matrix F	$dF = A\,dx$		

Ex: Infomax ICA

- Training examples $x_i \in \mathbb{R}^d$, $i = 1{:}n$
- Transformation $y_i = g(Wx_i)$
  - $W \in \mathbb{R}^{d \times d}$ parameter
  - $g(z) =$ scalar fn, componentwise
- Want: independent components



Ex: Infomax ICA

- $y_i = g(Wx_i)$
  - $dy_i = J(x_i, W)\,dx_i = J_i\,dx_i$

est. of $-\int P(y_i) \ln P(y_i)$

- Method: $\max_W \sum_i -\ln(P(y_i))$
  - where $P(y_i) = P(x_i) / |\det J(x_i, W)|$

$$\max_W \sum_i \left(\ln |\det J(x_i, W)| - \cancel{\ln P(x_i)} \right)$$