THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Statistics and Data Sciences
College of Natural Sciences

SDS 385: Stat Models for Big Data

Lecture 4a: starting with Ada-***

Purnamrita Sarkar
Department of Statistics and Data Science
The University of Texas at Austin

https://psarkar.github.io/teaching

Newton-Raphson

- We are optimizing $f(\beta)$
- Newton's method uses second order information :

$$\beta_{t+1} = \beta_t - [\nabla^2 f(\beta_t)]^{-1} \nabla f(\beta_t)$$

- Nice because

Newton-Raphson

- We are optimizing $f(\beta)$
- Newton's method uses second order information :

$$\beta_{t+1} = \beta_t - [\nabla^2 f(\beta_t)]^{-1} \nabla f(\beta_t)$$

- Nice because
  - Converges quicker than GD
  - stepsize is 1
- Problematic because?

Newton-Raphson

- We are optimizing $f(\beta)$
- Newton's method uses second order information :

$$\beta_{t+1} = \beta_t - [\nabla^2 f(\beta_t)]^{-1} \nabla f(\beta_t)$$

- Nice because
  - Converges quicker than GD
  - stepsize is 1
- Problematic because?
  - Computationally expensive when p is large
  - If you have a Nonconvex problem, Hessian does not have to be PSD.
  - If I am doing $\beta_{t+1} = \beta_t - \alpha G \nabla f(\beta_t)$ and G is PSD, then I claim the loss cannot increase for small α.

Natural gradient

- Consider a model where you are interested in getting the MLE.
- You are minimizing $-\log P(X;\beta)$
- The Hessian is $\nabla^2 - \log P(X;\beta)$
- What is the expectation of this quantity at β^*, such that $X \sim P(\cdot;\beta^*)$?

Natural gradient

- Consider a model where you are interested in getting the MLE.
- You are minimizing $-\log P(X;\beta)$
- The Hessian is $\nabla^2 - \log P(X;\beta)$
- What is the expectation of this quantity at β^*, such that $X \sim P(\cdot;\beta^*)$?
- $-E[\nabla^2 \log P(X;\beta^*)] = E[\nabla \log P(X;\beta^*) \nabla \log P(X;\beta^*)^T]$
- why?

Scalar case

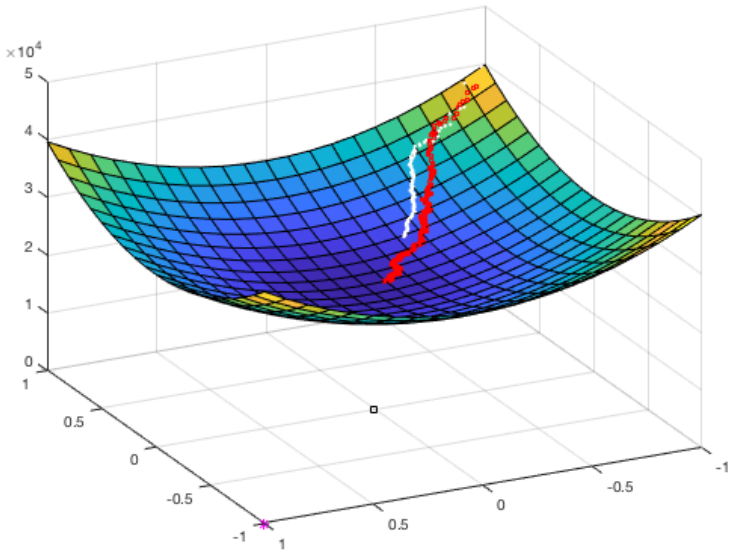
$$
\begin{aligned}
\nabla \log P(X;\beta^*) &= \frac{\nabla P(X;\beta^*)}{P(X;\beta^*)} \\
\nabla^2 \log P(X;\beta^*) &= \frac{P(X;\beta^*)\nabla^2 P(X;\beta^*) - \nabla P(X;\beta^*)^2}{P(X;\beta^*)^2} \\
&= \frac{\nabla^2 P(X;\beta^*)}{P(X;\beta^*)} - \frac{\nabla P(X;\beta^*)^2}{P(X;\beta^*)^2} \\
E\nabla^2 \log P(X;\beta^*) &= -\left(\nabla \log P(X;\beta^*)\right)^2
\end{aligned}
$$

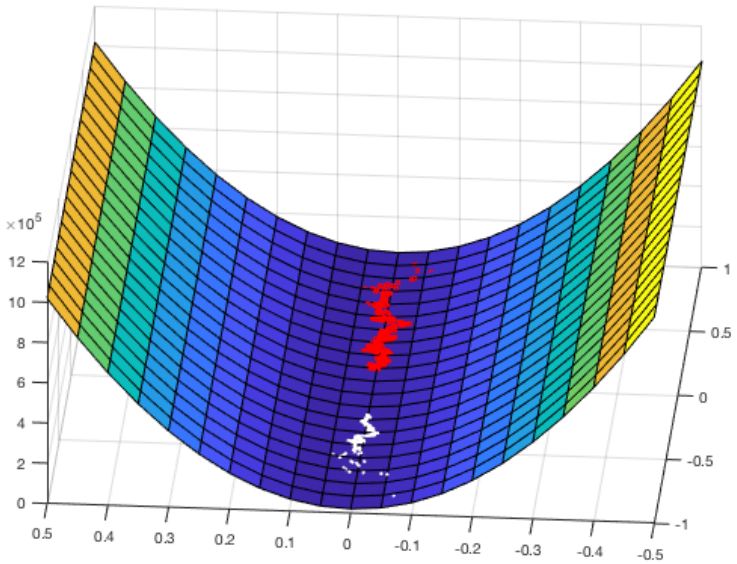
- Hope is that if we approximate the Hessian at w by $\sum_t g_t g_t^T / T$, where $g_t = \nabla \log P(x_t;\beta)$, then if $\beta_t \to \beta^*$, then the approximation is not too far off.

So where does Adagrad (Duchi, Hazan, Singer 2011) fit in?

- Use feature specific training rates.
- Intuition : sparse features are more informative.
- Use the previous gradients to obtain feature specific training rate.

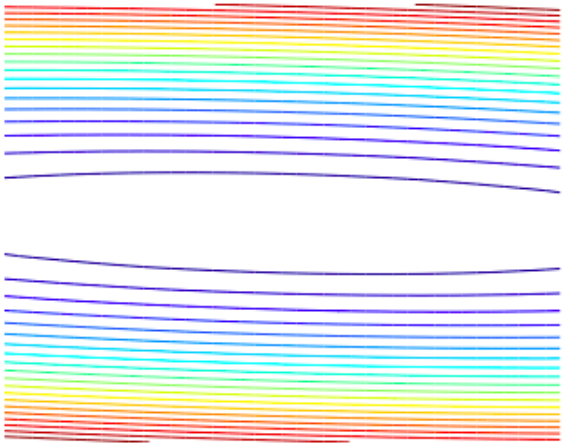
Adagrad: simple quadratic



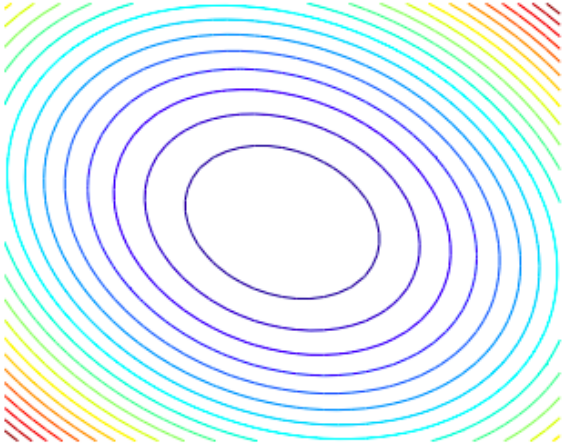


- The Hessian in the left panel is well conditioned, whereas in the second it is very skewed.
- Adagrad (red trajectory) seems to be progressing faster in the second.

A picture from Duchi et al's ISMP talk



Hard



Nice

y_t	$\phi_{t,1}$	$\phi_{t,2}$	$\phi_{t,3}$
1	1	0	0
-1	.5	0	1
1	-.5	1	0
-1	0	0	0
1	.5	0	0
-1	1	0	0
1	-1	1	0
-1	-.5	0	1

1. Frequent, irrelevant
2. Infrequent, predictive
3. Infrequent, predictive

Regret bound

- Standard regret bound: [We did this in class last time!]

$$\sum_{t=1}^{T} \left(f_t(\beta_t) - f_t(\beta^*) \right) \leq \frac{1}{2\alpha} \|\beta_0 - \beta^*\|^2 + \frac{\alpha}{2} \sum_t \|g_t\|^2$$

Regret bound

- Standard regret bound: [We did this in class last time!]

$$\sum_{t=1}^{T} \left(f_t(\beta_t) - f_t(\beta^*)\right) \leq \frac{1}{2\alpha}\|\beta_0 - \beta^*\|^2 + \frac{\alpha}{2}\sum_t \|g_t\|^2$$

- Regret bound that adapts to the geometry:

$$\sum_{t=1}^{T} \left(f_t(\beta_t) - f_t(\beta^*)\right) \leq \frac{1}{2\alpha}\|\beta_0 - \beta^*\|_A{}^2 + \frac{\alpha}{2}\sum_t g_t^T A^{-1} g_t$$

- Maholanobis distance: $\|x\|_A^2 = x^T A x$

In hindsight what A to choose?

- $\min_A \sum_t g_t^T A^{-1} g_t$, subject to A PSD, and trace of A is not too large.

In hindsight what A to choose?

- $\min_A \sum_t g_t^T A^{-1} g_t$, subject to A PSD, and trace of A is not too large.
- Solution: $A = C \left(\sum_t g_t g_t^T \right)^{1/2}$

In hindsight what A to choose?

- $\min_A \sum_t g_t^T A^{-1} g_t$, subject to A PSD, and trace of A is not too large.
- Solution: $A = C \left(\sum_t g_t g_t^T \right)^{1/2}$
- So at step t update $G_t = G_{t-1} + g_t g_t^T$
- Use $\beta_{t+1} = \beta_t - \alpha G_t^{-1/2} g_t$

Large dimensionality

- Cant do this for very large p
- Approximate with diagonal, aka

$$\beta_{t+1}(j) = \beta_t(j) - \alpha \frac{g_t(j)}{\sqrt{\sum_{i \leq t} g_i^2(j)}}$$

Large dimensionality

- Cant do this for very large p
- Approximate with diagonal, aka

$$\beta_{t+1}(j) = \beta_t(j) - \alpha \frac{g_t(j)}{\sqrt{\sum_{i \leq t} g_i^2(j)}}$$

- Each feature has its own rate.
- If a feature is rare, $\sum_i g_i^2(j)$ in general will be small, and you will weigh these more.

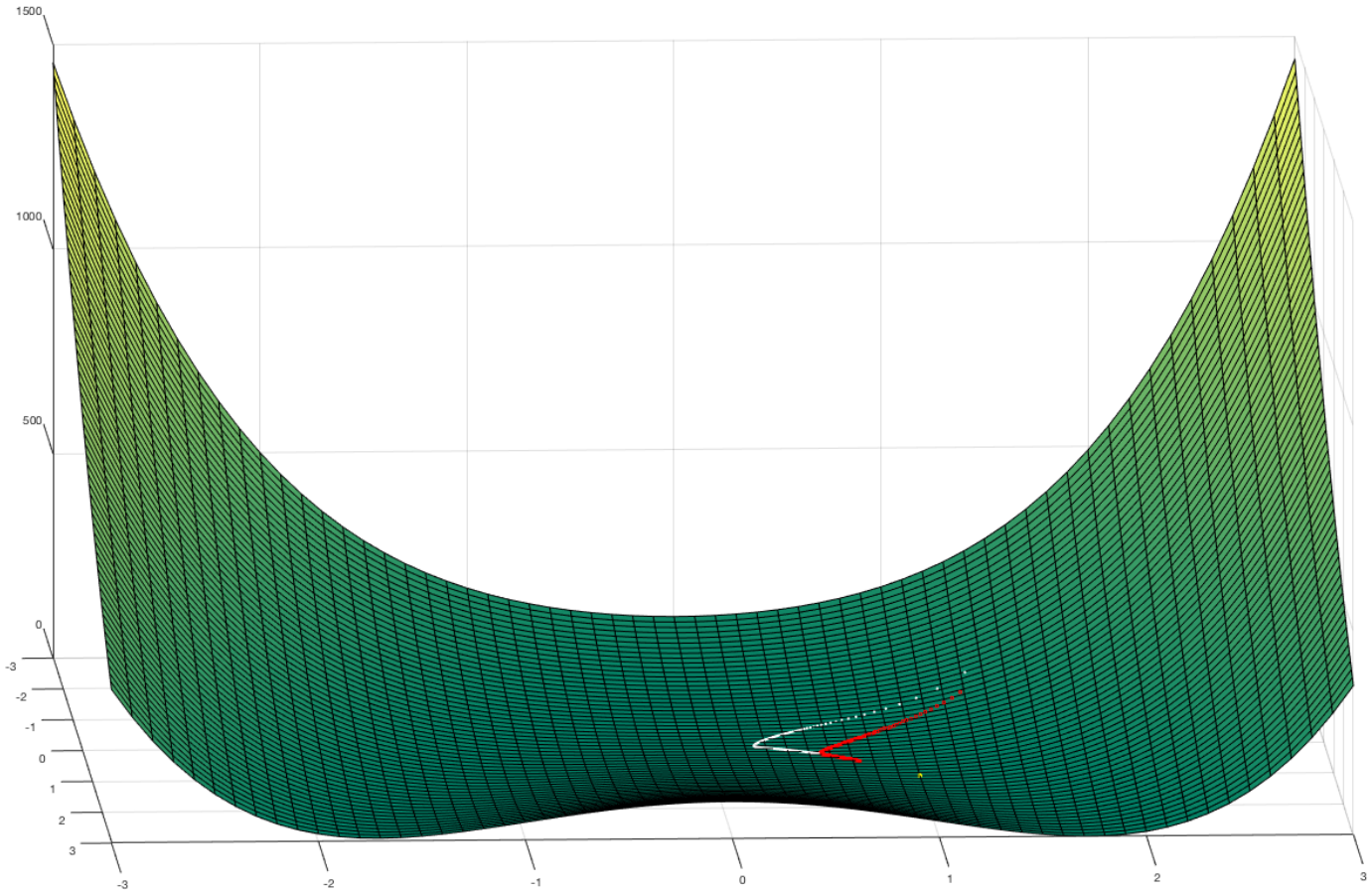
Final pretty picture

- The notorious Rosenbrock function. (see Wikipedia page if you want to know more).

$$f(x, y) = (a - x)^2 + b(y - x^2)^2$$

- Global minima: a, a^2.
- Quote from Wikipedia: The global minimum is inside a long, narrow, parabolic shaped flat valley. To find the valley is trivial. To converge to the global minimum, however, is difficult.

Final pretty picture



- Adagrad (red) finds the valley faster.
- Yellow dot is global optima.

ACK

- Duchi et al's paper
- Duchi et al's ISMP talk slides
- Sham Kakade's lecture notes