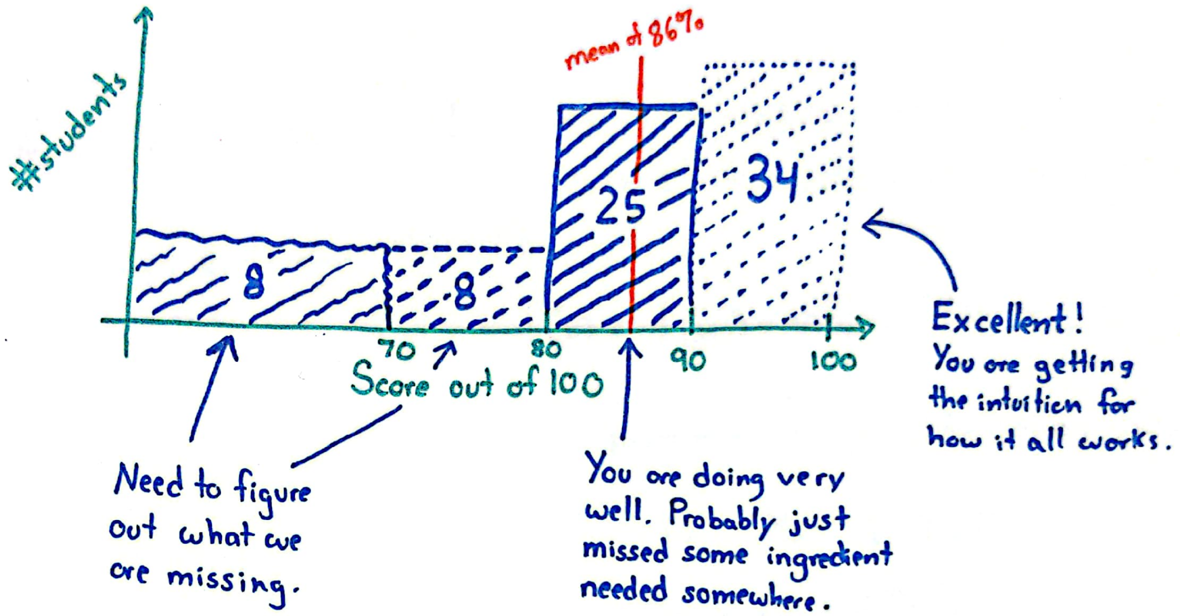
Midterms Graded / Posted

I'm happy with the general distribution.



The midterm can be worth anything from 15% - 40%
(whatever is best for you.)

Lots of room to pick things up going forward, if needed.

- (the final could be as high as 65%, so you can show improvement and be rewarded at any point.)
- (come by office hours with any uncertainties, its our only "social" learning setting, and I want reasons to give away 10% for participation)

When $\begin{cases} f_k = f(x_k) \\ g_k = \nabla f(x_k) \\ B_k = \nabla^2 f(x_k) \end{cases}$, this is just 2^{nd} order model used by Newton

If instead $B_k = LI$, we can recover the quadratic from our characterization of Lipz ∇f.

(L-Lipz grad $\Leftrightarrow$ $\nabla^2 f(x) \preceq LI$)

$$f(x) \approx f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{L}{2}\|x - x_k\|_2^2$$

The minimum/unique stationary point is $x_k - \frac{1}{L}\nabla f(x_k)$.

(recover GD).

$\boxed{\text{Starting 11/2}}$

Two corrections on HW4: (both changes have pushed to BB).

Q1 needs $\mathbb{E}\|g(x_k)\|^2 \leq M^2$

Q3(b) needs eta = 0.1 to avoid precision errors

(play with large eta values if you want to see how annoying parameter tuning can be.)

If $g_k = \frac{\partial f}{\partial x_i}(x) e_i$ instead (and keep $B_k = LI$), then we recover coordinate descent.

for $i \in \{1, \dots, d\}$

This motivates an algorithm like

$$x_{k+1} = \text{the stationary point of } m_k(x) \quad \text{or} \quad \underset{x}{\operatorname{argmin}} \{ m_k(x) \}$$

Well-defined if B_k is nonsingular

Well-defined if $B_k \succeq 0$. $\iff m_k$ is convex

$$\nabla m_k(x_k+p) = 0 \iff g_k + B_k p = 0$$

$$\iff p = -B_k^{-1} g_k$$

(recovers Newton step $-\nabla^2 f(x_k)^{-1} \nabla f(x_k)$)

Lets look at the geometry of a Newton step

$\nabla^2 f(x_k)$ is a symmetric, real matrix (and nonsingular)

$\Rightarrow$ Spectral decomposition

$$\nabla^2 f(x_k) = V \Lambda V^T \leftarrow \text{costs } O(d^3)$$

$$\Lambda = \text{diagonal matrix of eigenvalues} = \begin{pmatrix} \Lambda_+ & 0 \\ 0 & \Lambda_- \end{pmatrix} = \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_d \end{pmatrix}$$

$$V = \text{all of our corresponding eigenvectors} = \begin{pmatrix} V_+ & V_- \end{pmatrix} = \begin{pmatrix} v_1 & v_2 & \cdots & v_d \end{pmatrix}$$

The Newton step is then equal to

$$
\begin{aligned}
p^N &= -(V\Lambda V^T)^{-1}\nabla f(x_k) \\
&= -V\Lambda^{-1}V^T\nabla f(x_k) \qquad \text{(using } V \text{ is orthonormal)} \\
&= -(V_+ \; V_-)\begin{pmatrix}\Lambda_+^{-1} & 0 \\ 0 & \Lambda_-^{-1}\end{pmatrix}\begin{pmatrix}V_+^T \\ V_-^T\end{pmatrix}\nabla f(x_k) \\
&= -(V_+ \; V_-)\begin{pmatrix}\Lambda_+^{-1} & 0 \\ 0 & \Lambda_-^{-1}\end{pmatrix}\begin{pmatrix}V_+^T\nabla f(x_k) \\ V_-^T\nabla f(x_k)\end{pmatrix} \\
&= \underbrace{-V_+\Lambda_+^{-1}V_+^T\nabla f(x_k)}_{=:\, p_+^N} \underbrace{- V_-\Lambda_-^{-1}V_-^T\nabla f(x_k)}_{=:\, p_-^N}
\end{aligned}
$$

"Newton Step in the span of the positive eigenvectors".

"Newton Step in the span of the neg eigenvectors".

Claim: p_+^N is a descent direction (meaning $\underbrace{\nabla f(x_k)^T p_+^N}_{\text{directional derivative}} < 0$)

(Check: $-(\nabla f(x_k)^T V_+)\Lambda_+^{-1}(V_+^T\nabla f(x_k)) < 0$)

Symmetrically, p_-^N is an ascent direction ($\nabla f(x_k)^T p_-^N > 0$).

If all pos eigenvalues, descent

all neg " ", ascent

mixture of " ", could do anything

$(x_{k+1} = x_k + p)$

Lemma If $B_k > 0$, then $p_k = \text{argmin}\{\underbrace{g_k^T p + \frac{1}{2} p^T B_k p}\}$
has $g_k^T p_k < 0$.

In particular, if $g_k = \nabla f(x_k)$, this is a descent direction.

(needs B_k symmetric) otherwise $\frac{B_k + B_k^T}{2}$

Proof. Why is p_k well-defined? $B_k = \nabla^2 (g_k^T p + \frac{1}{2} p^T B_k p)$

$\succeq \lambda_{min}(B_k) I$

(since $B_k > 0$)

$\Rightarrow g_k^T p + \frac{1}{2} p^T B_k p$ is strongly convex.

$\Rightarrow$ Unique min exists.

$$
\begin{aligned}
g_k^T p_k &= g_k^T(-B_k^{-1} g_k) \\
&= -g_k^T \underbrace{B_k^{-1}}_{\text{p.d.}} g_k \\
&< 0. \quad \square
\end{aligned}
$$

(p_k has $\nabla m_k(p_k) = 0$ $\Leftrightarrow p_k = -B_k^{-1} g_k$)

It is not gouranteed that $x_{k+1} = \text{argmin}\{m_k(x)\}$ has $f(x_{k+1}) \leq f(x_k)$.

Only have $f(x_k + \alpha p_k) = f(x_k) + \alpha \underbrace{g_k^T p_k}_{<0} + o(\alpha^2)$.

Linesearch could be applied. Armijo condition for eta $\in (0,1)$

$$f(x_k - \alpha_k p_k) \leq f(x_k) + \eta \alpha_k g_k^T p_k$$

α_k exponentially shrinks until this found.

2. Modified Newton's Method

Given x_0

Iterate $k = 0, 1, 2, \ldots$

Compute $\nabla f(x_k)$, $\nabla^2 f(x_k)$

3 methods for this come → Build $B_k \succ 0$ (based on $\nabla^2 f(x_k)$)

Compute $p_k = \text{argmin} \{ \nabla f(x_k)^T p + \frac{1}{2} p^T B_k p \}$

$(= - B_k^{-1} p)$ (Solve $B_k p = -g_k$)

Pick α_k ensuring descent (Armijo linesearch)

HW5 constant choices work too

$$x_{k+1} = x_k + \alpha_k p_k$$

end loop.

$\boxed{\text{Option 1 for building } B_k}$ Throw away small / negative eigenvalues/vectors

Compute $\nabla^2 f(x_k) = V \Lambda V^T$

$\Lambda = \text{diag}(\lambda)$

Pick $\varepsilon > 0$ (reasonable choice for $\beta > 0$

$$\varepsilon = \begin{cases} \lambda_{max}(\nabla^2 f(x_k))/\beta & \text{if } \lambda_{max} > 0 \\ 1 & \text{otherwise} \end{cases}$$

condition # of B_k

$$\Rightarrow \frac{\lambda_{max}(B_k)}{\lambda_{min}(B_k)} = \frac{\lambda_{max}(\nabla^2)}{\varepsilon} = \beta.$$

$$\bar{\Lambda} = \text{diag}(\bar{\lambda}) \text{ where } \bar{\lambda}_i = \begin{cases} \lambda_i & \text{if } \lambda_i \geq \varepsilon \\ \varepsilon & \text{if } \lambda_i < \varepsilon. \end{cases}$$

$$B_k = V \bar{\Lambda} V^T \succ 0.$$

Previous lemma ensures $p_k = -B_k^{-1} \nabla f(x_k)$ descends.

If $\nabla f(x_k)$ has any component in negative or small eigen directions, $\|p_k\| \approx \frac{1}{\varepsilon}$.

mostly pointing in negative eigenvector directions.

Pretty bad direction unless $\nabla^2 f(x_k) \succeq \varepsilon I$, in which case Newton was good too.

$\boxed{\text{Option 2}}$ Move small values to ε, Keep large negative eigenvalues, but make them positive.

Compute $\nabla f(x_k)$, $\nabla^2 f(x_k) = V \Lambda V^T$

Pick $\varepsilon > 0$

$$\bar{\Lambda} = \mathrm{diag}(\bar{\lambda}), \quad \text{where } \bar{\lambda}_i = \begin{cases} \lambda_i & \text{if } \lambda_i \geq \varepsilon \\ \varepsilon & \text{if } -\varepsilon \leq \lambda_i \leq \varepsilon \\ -\lambda_i & \text{if } \lambda_i \leq -\varepsilon \end{cases}$$

$$B_k = V \bar{\Lambda} V^T \succ 0.$$