Recap. $\nabla f(x)$

$$f(x) = \frac{1}{2}\|Ax - y\|^2$$

$$\hookrightarrow \nabla f(x) = A^T(Ax - y)$$

$F: \mathbb{R}^p \to \mathbb{R}^q \qquad \underbrace{\partial F(x)}_{\mathbb{R}^p \xrightarrow[\text{linear}]{} \mathbb{R}^q} \in \mathbb{R}^{q \times p}$

$q = 1 \quad \nabla f(x) = \partial F(x)^T$

chain rule:

$$\partial(f \circ g) = \partial f \times \partial g$$

Ex. $f(x) = \ell(Bx)$

$f(x + \varepsilon\delta) + f(x) + \varepsilon \ldots$

$\to$ chain rule $f = \ell \circ B$

$$\nabla f(x) = B^T \cdot \nabla \ell(Bx)$$

$$\nabla f = B^T \circ \nabla \ell \circ B$$

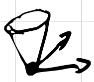
Exercise sheet 1.3 $\boxed{!}$

$$\|x\|_p = \left(\sum |x_i|^p\right)^{1/p}$$

$\|x + \varepsilon\delta\|_p$

$\hookrightarrow \|x\|_p = (g(x))^{1/p}$

Ex. $\|x\|_2 = \sqrt{x_1^2 + x_2^2}$



1.4

$$\langle X, Y \rangle_{\mathbb{R}^{n \times p}} = \sum_{ij} X_{ij} Y_{ij}$$

$$\langle X.\text{flatten}, Y.\text{flatten} \rangle_{\mathbb{R}^{np}}$$

$$\langle X, Y \rangle_{\mathbb{R}^{n \times p}} = \operatorname{trace}(XY^T) = \operatorname{tr}(YX^T) = \operatorname{tr}(X^TY)$$

$X, Y \in \mathbb{R}^{n \times p} \quad XY^T \in \mathbb{R}^{n \times n}$

$Z \in \mathbb{R}^{n \times n} \quad \operatorname{tr}(Z) = \sum Z_{ii} \qquad X^TY \in \mathbb{R}^{p \times p}$

① $\underbrace{(XY^T)_{ij}}_{n \times n} = \sum_k X_{ik} Y_{jk}$

$$\operatorname{tr}(XY^T) = \sum_i (XY^T)_{ii}$$
$$= \sum_i \sum_k X_{ik} Y_{ik}$$
$$= \sum_{ik} X_{ik} Y_{ik} = \langle X, Y \rangle$$

Application: $\langle AB, C \rangle_{\mathbb{R}^{n \times p}} = \langle B, A^TC \rangle$

$$\langle Ab, c \rangle = \langle b, A^Tc \rangle$$

Batch Gradient descent :

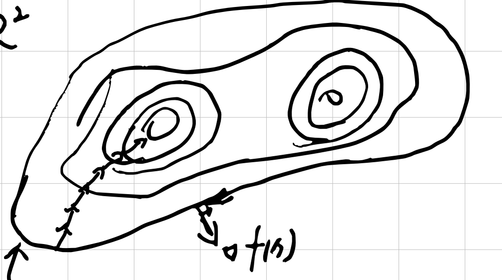
$\min_{x \in \mathbb{R}^d} f(x)$

$x_0 \leftarrow$ init

$x_{k+1} = x_k - \tau_k \nabla f(x_k)$

↳ learning rate > 0 ??

$x \in \mathbb{R}^2$



$\mathcal{L}_t = \{x : f(x) = t\}$

$\tau_k \to 0$
$\parallel$
τ

$$\frac{x_{k+1} - x_k}{\tau_k} = \nabla f(x_k)$$

$\tau \to 0 \downarrow$

$$x(t) = -\nabla f(x(t))$$

$\parallel$
$\frac{dx}{dt}$

↑ Gradient Flow ordinarity diff. Eq. (ODE)

Generic "descent" algorithm

$x_{k+1} = x_k - \tau_k d_k$

$f(x_{k+1}) = f(x_k) - \tau_k \langle d_k, \nabla f(x_k) \rangle + o(\tau_k)$

$\Delta_k = f(x_{k+1}) - f(x_k) \leq 0$ (< 0 if x_k is not minimum)

$\Delta_k = -\tau_k \langle d_k, \nabla f(x_k) \rangle + o(\tau_k)$

For $\Delta_k < 0$, for τ_k small enough

$\Leftrightarrow \langle d_k, \nabla f(x_k) \rangle > 0$

either $\nabla f(x_k) = 0$ $\boxed{\text{stop}}$

or $\nabla f(x_k) \neq 0$

$d_k = +\nabla f(x_k)$

↳ $\Delta_k = -\tau_k \|\nabla f(x_k)\|^2 + o(\tau_k)$

$d_k = H_k \cdot \nabla f(x_k)$

↳ sym & postive matrix

$H_k = U \cdot \mathrm{diag}(\lambda)\, U^T$

> 0 eigenvalues

$H_k = [\partial^2 f(x_k)]^{-1}$ Newton method

↑ convex

$\langle d_k, \nabla f(x_k) \rangle = \langle H_k \nabla f(x_k), \nabla f(x_k) \rangle$

$= \langle U \,\mathrm{diag}(\lambda) U^T \nabla f(x_k), \nabla f(x_k) \rangle$

$= \langle \mathrm{diag}(\sqrt{\lambda}) U^T \nabla f(x_k), \mathrm{diag}\sqrt{\lambda}\, U^T \nabla f(x_k) \rangle$

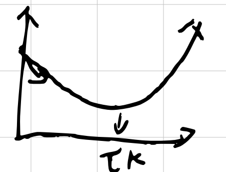
$= \|\mathrm{diag}(\sqrt{\lambda})\, U^T \nabla f(x_k)\|^2 > 0$

$f(x) = a x_1^2 + b x_2^2$ $\quad x^* = 0$

$a = b = 1$ $\quad a \ll b$

$h(\tau) = f(x_k - \tau \nabla f(x_k))$ $\quad h'(0) = -\|\nabla f(x_k)\|^2$

$\uparrow$
> 0



Armijo-Goldstein rule

numerical tours (website)

Theory $f(x) = \frac{1}{2}|Ax-y|^2 + \frac{\lambda}{2}|x|^2$

$f(x) = \frac{1}{2}\underbrace{\langle Cx, x\rangle}_{\text{quad}} - \underbrace{\langle x, b\rangle}_{\text{linear}}$

$C = A^TA + \lambda Id$

$b = A^Ty$

Ridge $f(x) = \frac{1}{2}|Ax-y|^2 + \frac{\lambda}{2}|x|^2$

$\nabla f(x) = A^T(Ax-y) + \lambda x$

General: $f(x) = \frac{1}{2}\langle Cx, x\rangle - \langle b, x\rangle$

$\nabla f(x) = Cx - b$

If $\ker(A) = \ker(C) = \{0\}$ (overdetermined)

$\boxed{A}$ then $x^* = C^{-1}b = (A^TA + \lambda Id)^{-1}(A^Ty)$

Operator norm / ∞-norm / algebra norm

$\|H\|_{op}$, $|||H|||$

Def: $\|H\|_{op} = \sup_{x \neq 0} \frac{\|Hx\|}{\|x\|}$

$\|Hx\|_2 \leq \|H\|_{op} \cdot \|x\|_2$

$\|H\|_{op}$ is the lipsch constant of

$x \mapsto Hx$
$\ell^2 \quad \ell^2$

$\|AB\|_{op} \leq \|A\|_{op} \cdot \|B\|_{op}$

Prop: $\|H\|_{op} = \sup_i \underbrace{\sqrt{\underbrace{\lambda_i}_{\text{eigenvalue}}(H^TH)}}_{\text{singleton value}}$

If H is symmetry, $\|H\|_{op} = \sup_i |\lambda_i(H)|$

Proof: $\|Hx\|^2 = \langle Hx, Hx\rangle = \langle \underbrace{H^TH}_{\text{sym} \geq 0} x, x\rangle$

$(H^TH = U \operatorname{diag}(\lambda) U^T)$

$= \langle U \operatorname{diag}(\lambda) U^T x, x\rangle$

$= \langle \operatorname{diag}(\lambda) \underbrace{U^Tx}_{z}, \underbrace{U^Tx}_{z}\rangle$

$= \sum_i \lambda_i z_i^2$

$\leq \max(\lambda) \underbrace{\sum_i z_i^2}_{\|U^Tx\|^2 = \|x\|^2}$

$\|Hx\|^2 \leq \max(\lambda)\|x\|^2$

$\|Hx\| \leq \sqrt{\max \lambda_i (H^TH)}\, \|x\|$

$U = (u_1, u_2 \dots u_n)$

$\lambda_1 \geq \lambda_2 \dots \geq \lambda_n$

$x = u_1$

$\|Hu_1\| = \sqrt{\max(\lambda)}\, \|u_1\|$

$= \|H\|_{op}$

Theory of GD $\rightarrow$ strongly cvx
C is invertible **FAST**
over determined

$\searrow$ C might be non-inv **slow**
under-determined

"Nice" case : C is inv

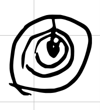
$$0<\lambda_1(C)\le\lambda_2(C)\le\cdots\le\lambda_d(C)$$

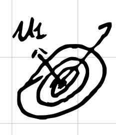
$$\overset{\shortparallel}{\mu} \qquad\qquad L=\|C\|_{op}$$

$$K=\frac{L}{\mu}\ge 1 \quad \text{conditioning}$$

$$K=1 \quad \lambda_1=\lambda_2=\cdots=\lambda_d \quad C=\lambda Id$$

$$K\gg 1 \quad u_1 \quad u_2$$





$$f(x)=\frac{1}{2}\langle Cx,x\rangle-\langle b,x\rangle$$

$$\nabla f(x)=Cx-b \qquad x^*=C^{-1}b$$

$$0=Cx^*-b$$

Fixed LR $\tau_k=\tau$ $\qquad x_{k+1}=x_k-\tau\nabla f(x_k)$

$$-\begin{cases} x_{k+1}=x_k-\tau(Cx_k-b) & \Delta_k=x_k-x^* \\ x^*=x^*-\tau(\underbrace{Cx^*-b}_{=0}) \end{cases}$$

$$\Delta_{k+1}=\Delta_k-\tau C\Delta_k$$

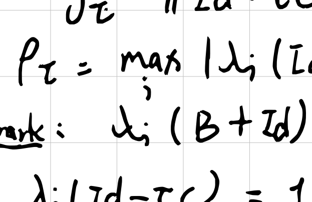
$$\Delta_{k+1}=(Id-\tau C)\Delta_k$$

$$\Delta_k=(Id-\tau C)^k\Delta_0$$

$$\|\Delta_k\|\le\underbrace{\|Id-\tau C\|}_{=\rho}{}^k\underbrace{\|\Delta_0\|}_{\|x_0-x^*\|}$$

$\rho>1\rightarrow$ esplodes $\rightarrow$ no convergence

$\rho<1\rightarrow$ "linear convergence" ρ^k



$\rightarrow$

$$\|\Delta_k\|\le\rho^k\|\Delta_0\| \rightarrow \log(\|\Delta_k\|)\le k\log(\rho)+cst$$

Summary :

$$\rho_\tau=\|Id-\tau C\|_{op}$$

$$\rho_\tau=\max_i|\lambda_i(Id-\tau C)|$$

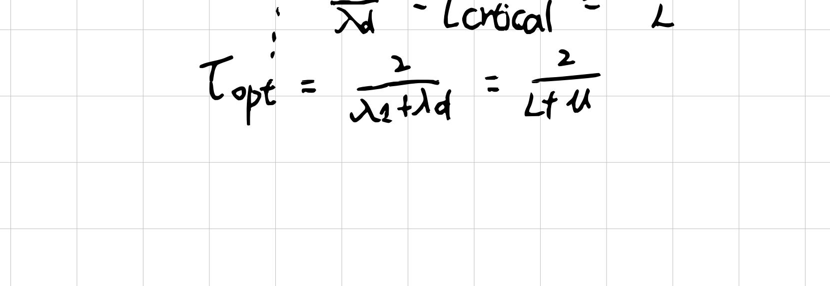
Remark: $\lambda_i(B+Id)=\lambda_i(B)+1$

$$\lambda_i(Id-\tau C)=1-\tau\lambda_i(C)$$

$$\rho_\tau=\max|1-\tau\lambda_1(C)|,|1-\tau\lambda_2(C)|,|1-\tau\lambda_d(C)|$$

$$\lambda_1\le\lambda_2\le\cdots$$

$$1-\tau\lambda_1=0 \qquad \tau=\frac{1}{\lambda_1}$$



$$\frac{2}{\lambda_d}=\tau_{critical}=\frac{2}{L}$$

$$\tau_{opt}=\frac{2}{\lambda_1+\lambda_d}=\frac{2}{L+\mu}$$

Thm : If $0<\mu=\lambda_1(C)\le\lambda_2(C)\le\cdots\le\lambda_d(C)=L$
If $0<\tau<\frac{2}{L}$, then $\exists\rho<1$
s.t. $\|x_k-x^*\|\le\rho^k\|x_0-x^*\|$
A "good" (optimal for the proof)
choice is $\tau=\frac{2}{L+\mu}$

$$\rho_{\tau_{op}}=\frac{L-\mu}{L+\mu}=\frac{K-1}{K+1}$$

$K=\frac{L}{\mu}$ $\qquad K\to+\infty \quad \rho_{\tau_{op}}\to 1$

$$K\to 1 \quad \rho_{\tau_{op}}\to 0$$

Underdetermined $\mu=0$, $K=+\infty$

Thm : If $0<\tau_{min}\le\tau_k\le\frac{2}{L}$
u_k will converge to same sol° x^*

$$f(x_k)-f(x^*)\le\frac{dist(x_0,\text{argmin}(f))}{8\times\tau\times k}$$

$\sim$ true e^2? 2 3 4

General case : $C\longrightarrow$ Hessian

Def: $f:\mathbb{R}^d\to\mathbb{R}$

$$\partial^2 f(x)=\left(\frac{\partial f}{\partial x_i x_j}(x)\right)_{i,j=1}^d\in\mathbb{R}^{d\times d}$$

$$\partial^2 f(x)^T=\partial^2 f(x)$$

Prop: f is cvx $\Leftrightarrow\partial^2 f(x)\succeq 0$

$$\lambda_i(\partial f^2(x))\ge 0$$

Def: f is twice differtiable at x

$$f(x+\varepsilon\delta)=\underbrace{f(x)}_{cst}+\underbrace{\varepsilon\langle\nabla f(x),\delta\rangle}_{linear}+\underbrace{\frac{\varepsilon^2}{2}\langle\partial^2 f(x)\times\delta,\delta\rangle}_{quad}+o(\varepsilon^2)$$

$$\hookrightarrow\sum_{ij}\frac{\partial^2 f}{\partial u_i\partial u_j}(x)\times\delta_i\times\delta_j$$

$$\nabla f(x+\varepsilon\delta)=\nabla f(x)+\varepsilon\partial^2 f(x)\times\delta+o(\varepsilon)$$

Thm 1 : $0<\mu\le\lambda_i(\partial^2 f(x))\le L$ $\quad f$ is convex

μ: strong cvxity $\qquad L$: smoothness $\qquad \Updownarrow\ \mu\ge 0$

Remark : if $f(x)=\frac{1}{2}\langle Cx,x\rangle-\langle b,x\rangle$

$$\nabla f(x)=Cu-b$$

$$\partial^2 f(x)=C$$

$$\mu=\inf_x\lambda_i(\partial^2 f(x))$$

$$L=\sup_x\lambda_i(\partial^2 f(x)) \qquad L\ne\lambda_{max}(\partial^2 f(x))$$

Same thm : $0<\tau<\frac{2}{L}\Rightarrow$ conv

$$\|x_k-x^*\|\le\rho^k\|x_k-x_0\|$$

Thm2 : same $\mu=0$, $\tau<\frac{2}{L}$

$$x_k\to x^k$$

$$f(x_k)-f(x^*)\le O(1/k)$$

$$\partial^2 f(x)\le L\,Id$$

$$\mu\le\lambda_i(\partial^2 f(x))\le L$$

$$[f(x_k)+\langle\nabla f(x_k),x-x_k\rangle+\frac{\mu}{2}\|x-x_k\|^2]$$

$$\le f(x)$$

$$\le[f(x_k)+\langle\nabla f(x_k),x-x_k\rangle+\frac{L}{2}\|x-x_k\|^2]$$

GD $\sim 1/k$

Vesterov $\sim 1/k^2$ optimal