Optimization

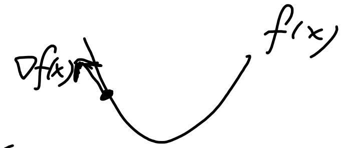
$\min\limits_{x\in\mathbb{R}^d} f(x)$ or $\min\limits_{x\in\mathcal{D}\subset\mathbb{R}^d} f(x)$

$x^* = \arg\min\limits_{x\in\mathcal{D}\subset\mathbb{R}^d} f(x)$

"Gradient-based optimization"

✓ $\langle \underbrace{\nabla f(x)}, \vec{l} \rangle = \lim\limits_{\varepsilon\to 0} \dfrac{f(x+\varepsilon\vec{l}) - f(x)}{\varepsilon}$ (☆)

✓ $\nabla f(x) = \left(\dfrac{\partial f}{\partial x_1}, \dots, \dfrac{\partial f}{\partial x_d}\right)$

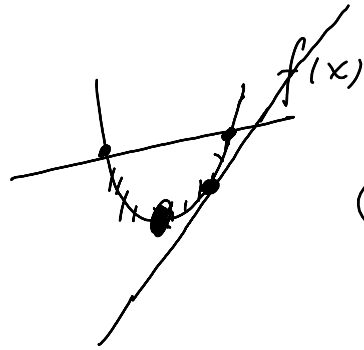


$x^0 \xrightarrow{\checkmark} x^1 \xrightarrow{\checkmark} x^2 \xrightarrow{\checkmark} \cdots \xrightarrow{\checkmark} x^{(N)} \approx x^*$

$x^k \xrightarrow[\nabla^2 f(x)]{\nabla f(x)} x^{k+1}$

$\vdots$

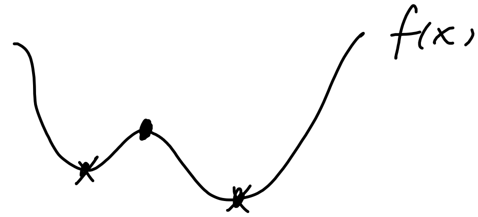
Convex function

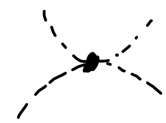


Convex optimization

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

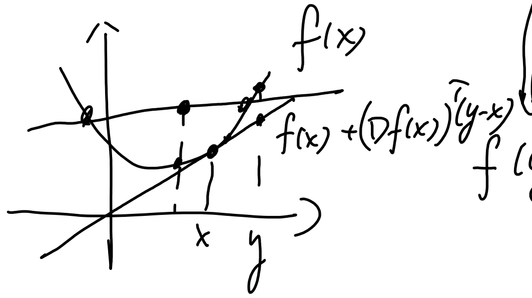
Non-Convex optimization





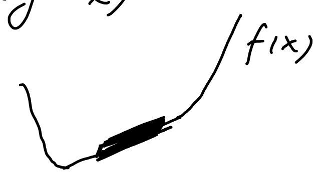
f is convex $\quad f: \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$.

$\forall \alpha \in [0,1] \quad f((1-\alpha)x+\alpha y) \underset{\text{"="}}{\le} (1-\alpha) f(x) + \alpha f(y)$



$\Updownarrow$ say f is convex and ∇f is continuous

$$f(y) \ge f(x) + (\nabla f(x))^T (y-x)$$



Strongly Convex Functions $\alpha \in [0,1]$ $\underline{m > 0}$ convexity constant

$$f((1-\alpha)x + \alpha y) \leq (1-\alpha) f(x) + \alpha f(y) - \frac{1}{2} m \alpha (1-\alpha) \|x-y\|_2^2$$

$m = 0$ convex functions

Wednesday, July 22, 2020 8:55 AM

Taylor's Theorem

$f: \mathbb{R}^n \to \mathbb{R} \quad x, p \in \mathbb{R}^n$

$$\boxed{f(\underbrace{x+p}_{y}) = f(x) + \int_0^1 \nabla f(x+\gamma p)^T p \, d\gamma}$$

$p = y - x$

$g(\gamma) = f(x + \gamma p)$ $\quad g(1) - g(0) = \int_0^1 g'(\gamma) d\gamma = g'(\tilde{\gamma})$

$\gamma \in [0,1]$ $\quad = \int_0^1 \nabla f(x+\gamma p)^T p \, d\gamma$

$\gamma p = (\gamma p_1, \ldots, \gamma p_n)$

$f(x+p) = f(x) + \nabla f(x + \tilde{\gamma} p)^T p$ for some $\tilde{\gamma} \in [0,1]$

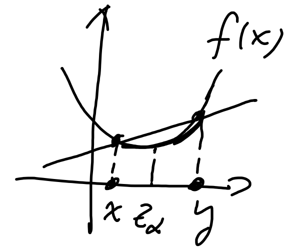
$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+\gamma p)^T p \, d\gamma$$

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ - & - & & - \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

- $$\boxed{f(x+p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 \underline{f(x+\gamma p)}\, p.}$$

for some $\gamma \in [0,1]$

Convex function



- $f((1-\alpha)x+\alpha y) \le (1-\alpha)f(x)+\alpha f(y)$

$z_\alpha = (1-\alpha)x+\alpha y$

if $\alpha \to 0$ then $z_\alpha \to x$

if $\alpha \to 1$ then $z_\alpha \to y$

$z_\alpha - x$
$= \alpha(y-x)$
$y > x$

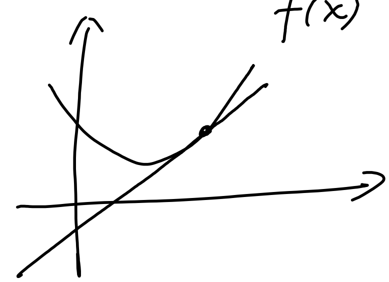
$$f(z_\alpha) \le (1-\alpha)f(x)+\alpha f(y)$$

$$\Leftrightarrow \quad \alpha f(y) - \alpha f(x) \ge f(z_\alpha) - f(x)$$

$$\Leftrightarrow \quad \frac{\cancel{\alpha} f(y) - \cancel{\alpha} f(x)}{\cancel{\alpha}(y-x)} \ge \frac{f(z_\alpha)-f(x)}{z_\alpha - x} \qquad \alpha \downarrow 0$$

$$\frac{f(y)-f(x)}{y-x} \geqslant \lim_{\alpha \downarrow 0} \frac{f(z_\alpha)-f(x)}{z_\alpha - x} = f'(x)$$

$$f(y)-f(x) \geqslant f'(x)(y-x)$$



$$f(y)-f(x) \geqslant (\nabla f(x))^T (y-x)$$

$$f(y) \geqslant f(x) + (\nabla f(x))^T (y-x)$$

Strongly convex function

$$f(\underbrace{(1-\alpha)x+\alpha y}) \le (1-\alpha)f(x)+\alpha f(y) - \frac{m}{2}\alpha(1-\alpha)\|x-y\|_2^2$$

$$z_\alpha = (1-\alpha)x+\alpha y \qquad z_\alpha - x = \alpha(y-x)$$

$$\cancel{\alpha} f(y) - \cancel{\alpha} f(x) \geq f(z_\alpha) - f(x) + \frac{1}{2}m\alpha(1-\alpha)\|x-y\|_2^2$$

$$= (\nabla f(x))^T \cancel{\alpha}(y-x) + \cancel{O(\alpha^2\|y-x\|^2)}$$

$$+ \frac{1}{2}m\cancel{\alpha}\underbrace{(1-\alpha)}_{1}\|x-y\|_2^2$$

$$\alpha \to 0 \qquad f(y)-f(x) \geq (\nabla f(x))^T(y-x) + \frac{1}{2}m\|x-y\|^2$$

- $f(y) \geq f(x) + (\nabla f(x))^T (y-x) + \frac{1}{2} \underline{m} \|x-y\|_2^2$

$$\nabla^2 f(x) \succeq mI$$

$$\lambda(\nabla^2 f(x)) \geq m$$

$$f(x) \geq f(z) + (\nabla f(z))^T (x-z) + \frac{m}{2} \|x-z\|^2 \quad ①$$

$$f(y) \geq f(z) + (\nabla f(z))^T (y-z) + \frac{m}{2} \|y-z\|^2 \quad ②$$

$(1-\alpha) \times ① + \alpha \times ②$

$$(1-\alpha) f(x) + \alpha f(y) \geqslant f(z) + \cancel{(\nabla f(z))^T [(1-\alpha)(x-z) + \alpha (y-z)]}$$

$$+ \frac{m}{2} \left[(1-\alpha) \|x-z\|^2 + \alpha \|y-z\|^2 \right]$$

$$\underset{\sim}{z} = z_\alpha = \underline{(1-\alpha) x + \alpha y}$$

$$z - x = -\alpha (x-y)$$

$$z - y = (1-\alpha) x + (\alpha - 1) y$$

$$= (1-\alpha)(x-y)$$

$$(1-\alpha)(x-z) + \alpha (y-z) = (1-\alpha) \alpha \cancel{(x-y)} + \alpha (1-\alpha) \cancel{(y-x)} = 0$$

$$(1-\alpha) \|x-z\|^2 + \alpha \|y-z\|^2 = \alpha (1-\alpha) \|x-y\|^2$$

$$(1-\alpha) f(x) + \alpha f(y) \geq f(\alpha x + (1-\alpha) y) + \frac{m}{2} \alpha (1-\alpha) \|x-y\|^2$$

$$f(\alpha x + (1-\alpha) y) \leq (1-\alpha) f(x) + \alpha f(y) - \frac{m}{2} \alpha (1-\alpha) \|x-y\|^2$$

Optimization

Assumptions ① Df is L-Lipschitz

$$\|Df(x) - Df(y)\| \leq L \|x-y\|$$

② f is m-strongly convex

$$(1-\alpha) f(x) + \alpha f(y) - \frac{m}{2} \alpha (1-\alpha) \|x-y\|^2 \geq f(\alpha x + (1-\alpha) y).$$

$$f(x)+(\nabla f(x))^T(y-x)+\frac{m}{2}\|y-x\|^2 \leq f(y) \leq f(x)+(\nabla f(x))^T(y-x)+\frac{L}{2}\|y-x\|^2$$

$$\boxed{LI \geq \nabla^2 f(x) \geq mI}$$

$$\lambda_{max}(\nabla^2 f(x)) \leq L$$

$$\lambda_{min}(\nabla^2 f(x)) \geq m$$

$$f(y) = f(x) + \int_0^1 \nabla f(x+\gamma(y-x))^T(y-x)\,d\gamma$$

$$f(y)-f(x)-(\nabla f(x))^T(y-x) = \int_0^1 \left[\nabla f(x+\gamma(y-x))^T-(\nabla f(x))^T\right](y-x)\,d\gamma$$

$$a^Tb \leq \|a\|\cdot\|b\|$$

$$\leq \int_0^1 \|\nabla f(x+\gamma(y-x))-\nabla f(x)\|\cdot\|y-x\|\,d\gamma$$

$$\langle a,b\rangle \leq \|a\|\cdot\|b\|$$

$$\leq \int_0^1 L\underline{\gamma}\|y-x\|^2\,d\gamma = \frac{L}{2}\|y-x\|^2$$

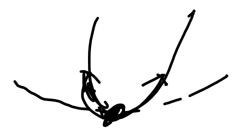
$$\nabla f(x) = Qx + b \qquad \nabla^2 f(x) = Q$$

$$f(x) = \frac{1}{2} x^T Q x + b^T x + c \qquad \frac{1}{2}(x_1 \dots x_d)\begin{pmatrix} q_{11} & \cdots & q_{1d} \\ & \cdots & \\ q_{d1} & \cdots & q_{dd} \end{pmatrix}\begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$$

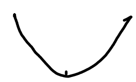
$$= \frac{1}{2}\sum_{i,j} q_{ij} x_i x_j .$$

$$\langle \nabla f(x), \vec{\ell} \rangle$$

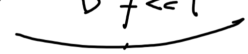
$$= \lim_{\varepsilon \to 0} \frac{\frac{1}{2}(x+\varepsilon\vec{\ell})^T Q (x+\varepsilon\vec{\ell}) + b^T(\cancel{x}+\varepsilon\vec{\ell}) + \cancel{c} - \left(\cancel{\frac{1}{2}x^T Q x} + \cancel{b^T x} + \cancel{c}\right)}{\varepsilon}$$



$$= \lim_{\varepsilon \to 0} \frac{\frac{1}{2}\cancel{\varepsilon}\, \ell^T Q x + \frac{1}{2}\cancel{\varepsilon}\, x^T Q \vec{\ell} + \cancel{\varepsilon}\, b^T \vec{\ell}}{\cancel{\varepsilon}} = \langle Qx + b, \vec{\ell} \rangle$$



$\nabla^2 f >> 1$

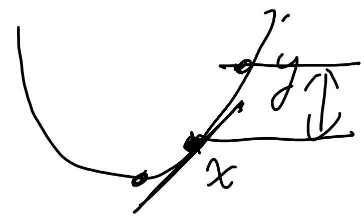


$\nabla^2 f << 1$

Wednesday, July 22, 2020 9:43 AM

If f is strongly convex with convexity constant $m > 0$

$$f(y) - f(x) \geq -\frac{1}{2m} \|\nabla f(x)\|^2$$



$$\boxed{\|\nabla f(x)\|^2 \geq 2m\,[f(x) - f(y)]}$$

$$f(x^k) \to f(x^*) \iff \nabla f(x^k) \to 0$$

$$\|\nabla f(x^k)\|^2 \geq 2m\,[f(x^k) - f(x^*)]$$

$$f(y) \geq f(x) + (\nabla f(x))^T (y-x) + \frac{m}{2} \|y-x\|^2$$

$$= f(x) + \frac{m}{2}\left[\|y-x\|^2 + \frac{2}{m}(\nabla f(x))^T (y-x)\right]$$

$$= f(x) + \frac{m}{2}\left[\left\|\underline{y - x + \frac{1}{m}\nabla f(x)}\right\|^2 - \frac{1}{m^2}\|\nabla f(x)\|^2\right]$$

$$\geq f(x) - \frac{1}{2m}\|\nabla f(x)\|^2$$

Monday, July 27, 2020 8:32 AM

GD AGD

"first-order algorithms"

Optimization Algorithm

$$\underbrace{\min_{x \in \mathbb{R}^n}} f(x) = f(x^*)$$

$$x^0 \to x^1 \to x^2 \to \cdots \to x^k \to \cdots$$

$$x^k \to x^* \quad \text{as } k \to \infty$$

$$x^k \xrightarrow{?} \underline{x^{k+1}}$$

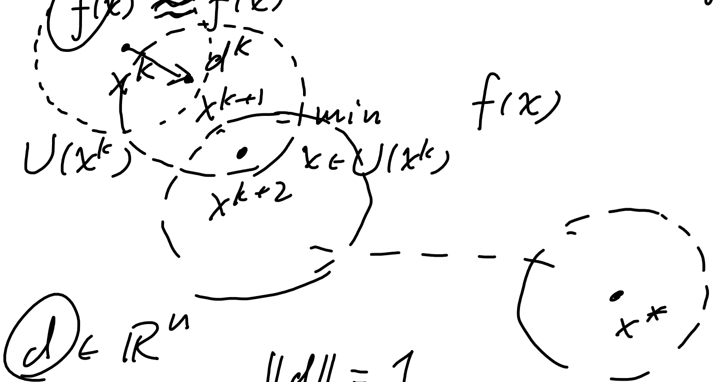
$f(x^k)+\nabla f(x^k)^T(x-x^k)\cancel{+\frac{1}{2}(x-x^k)^T\nabla^2 f(x^k)(x-x^k)}$

$= \hat{f}(x) \approx f(x)$

Local neighborhood

"trust region"

Line Search Methods



For $x \in \mathbb{R}^n$ and $d \in \mathbb{R}^n$ $\quad \underline{\|d\| = 1}$

descent direction

$$f(x + t d) < f(x) \quad \text{for some } \underline{t \text{ small}}$$

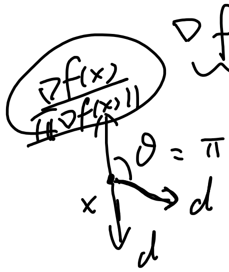
$$f(x+td) = f(x) + t\,\underline{\nabla f(x+\gamma t d)^T d} \quad \text{for some } \gamma \in (0,1)$$

$$< f(x)$$

$$\nabla f(x+\underline{\gamma t d})^T d < 0 \qquad \inf_{\|d\|=1} (\nabla f(x))^T d$$

ss

$$\nabla f(x)^T d \qquad = \inf_{\|d\|=1} \|\nabla f(x)\| \cdot \left\langle \frac{\nabla f(x)}{\|\nabla f(x)\|}, d \right\rangle$$



$$\downarrow$$

$$d = -\frac{\nabla f(x)}{\|\nabla f(x)\|} \qquad = \|\nabla f(x)\| \underline{\inf_{\|d\|=1} \left\langle \frac{\nabla f(x)}{\|\nabla f(x)\|}, d \right\rangle}$$

$$x \longrightarrow x + t \cdot \left(\frac{-\nabla f(x)}{\|\nabla f(x)\|} \right)$$

$$= x - \underbrace{\boxed{\frac{t}{\|\nabla f(x)\|}}}_{\alpha(x)} \nabla f(x)$$

$$\boxed{x^{k+1} = x^{k} - \underbrace{\alpha_k}_{\text{learning rate}} \nabla f(x^{k})}$$

Gradient Descent

Proof of Convergence

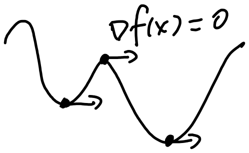
Assume f is m-strongly convex, ∇f is L-Lipschitz

$$mI \preceq \nabla^2 f(x) \preceq LI$$

$$m \le \lambda(\nabla^2 f(x)) \le L$$

① **Spectral Method**

$$x^{k+1} = x^k - \alpha \nabla f(x^k)$$
$$= \Phi_\alpha(x^k)$$



$$x = \phi(x)$$

$$\Phi_\alpha(x) = x - \alpha \nabla f(x) \rightsquigarrow \iff \nabla f(x) = 0$$

$$x^{k+1} = \Phi_{\alpha_k}(x^k) \qquad x^k \longrightarrow \underline{\underline{x^*}} \qquad \underbrace{x^* = \Phi_\alpha(x^*)}_{\text{fixed pt.}}$$

$$x^k = \Phi_\alpha(x^{k-1}) = \Phi_\alpha \circ \Phi_\alpha(x^{k-2})$$

$$= \cdots = \underbrace{\left(\Phi_\alpha \circ \Phi_\alpha \circ \cdots \circ \Phi_\alpha\right)}_{k}(x^0)$$

"contraction mapping theorem" $\boxed{\|\Phi(x) - \Phi(y)\| \le \beta\|x-y\|}$ $\underline{0<\beta<1}$

$$\|x^{k+1} - x^*\| = \|\Phi(x^k) - \Phi(x^*)\| \le \beta\|x^k - x^*\|$$

$$\le \beta^2\|x^{k-1} - x^*\| \le \cdots \le \beta^k\|x^0 - x^*\|$$

$$x^{k+1} \longrightarrow x^* \qquad \longrightarrow 0 \ (k\to\infty)$$

$$\|\phi(x)-\phi(y)\| = \| x-\alpha Df(x) - (y-\alpha Df(y))\|$$

$$= \|(x-y) - \alpha\,(Df(x) - Df(y))\|$$

$$= \left\|(x-y) - \alpha\int_0^1 \underbrace{D^2 f(y+\gamma(x-y))}_{\in[m,L]}\,\underline{(x-y)}\,d\gamma\right\|$$

$$= \left\|\underbrace{\left(I - \alpha\int_0^1 \underline{D^2 f(y+\gamma(x-y))}\,d\gamma\right)}\underline{\underline{(x-y)}}\right\|$$

$$1-\alpha L \le \lambda(A(x,y)) \le 1-\alpha m$$

Monday, July 27, 2020 9:07 AM

$$\| \phi(x) - \phi(y) \| = \| \underline{A(x, y)} (x-y) \|$$

$$\leq \beta \| x-y \| \quad 0 < \beta < 1$$

$|\lambda(A)| \leq \beta$ " spectral method"

$$-\beta \leq \underbrace{1-\alpha L}_{\alpha \leq \frac{1+\beta}{L}} \leq \lambda(A) \leq 1-\alpha \underbrace{m \leq \beta}_{\alpha \geq \frac{1-\beta}{m}}$$

$$\frac{1-\beta}{m} = \frac{1+\beta}{L} \Leftrightarrow \beta = \frac{L-m}{L+m}$$

$$x^{k+1} = x^k - \alpha \nabla f(x^k) \qquad \alpha \in \left[\frac{1-\beta}{m}, \frac{1+\beta}{L}\right]$$

$$\alpha = \frac{1-\frac{L-m}{L+m}}{m} = \frac{2}{L+m}$$

② Taylor Expansion Method

"Generally"

$$f(x+\alpha d) = f(x) + \alpha(\nabla f(x))^T d + \alpha \int_0^1 \underbrace{[\nabla f(x+\gamma\alpha d) - \nabla f(x)]^T d\, d\gamma}_{\langle \nabla f(x+\gamma\alpha d) - \nabla f(x),\, d\rangle}$$

$$\leq f(x) + \alpha(\nabla f(x))^T d + \alpha \int_0^1 \|\nabla f(x+\gamma\alpha d) - \nabla f(x)\| \cdot \|d\|\, d\gamma$$

"∇f is L-Lipschitz"

$$\leq f(x) + \alpha(\nabla f(x))^T d + \alpha \int_0^1 L\gamma\alpha \|d\|^2 d\gamma$$

$$= f(x) + \alpha(\nabla f(x))^T d + \frac{L\alpha^2\|d\|^2}{2}$$

"if $\alpha = \frac{1}{L}$ then $\alpha - \frac{\alpha^2 L}{2} = \frac{1}{L} - \frac{\frac{1}{L}}{2} = \frac{1}{2L}$"

Monday, July 27, 2020 9:19 AM

$$f(x+\alpha d) \le f(x) + \underbrace{\alpha (\nabla f(x))^T d + \frac{\alpha^2 L}{2} \|d\|^2}$$

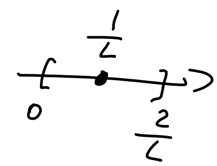
$$< f(x)$$

GD $\quad x^{k+1} = x^k - \alpha \nabla f(x^k) \qquad d = -\nabla f(x)$

$$f(x^{k+1}) \le f(x^k) + \alpha (\nabla f(x^k))^T (-\nabla f(x^k)) + \frac{\alpha^2 L}{2} \|\nabla f(x^k)\|^2$$

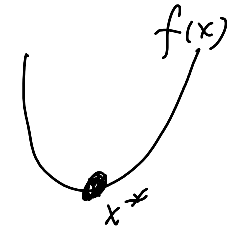
$$= f(x^k) - \underline{\underline{\left(\alpha - \frac{\alpha^2 L}{2}\right)}} \|\nabla f(x^k)\|^2$$

$0 < \alpha \left[1 - \frac{\alpha L}{2}\right] \iff \alpha < \frac{2}{L}$ $\qquad \alpha \in \left(0, \frac{2}{L}\right)$ $\quad f(x^{k+1}) < f(x^k)$



$\alpha = \frac{1}{L}$

$$x^{k+1} = x^k - \frac{1}{L} \nabla f(x^k)$$



$$f(x^{k+1}) \overset{①}{\le} f(x^k) - \frac{1}{2L} \| \nabla f(x^k) \|^2$$

If f is m-strongly convex $\quad \| \nabla f(x^k) \|^2 \overset{②}{\geq} 2m [f(x^k) - f(x^*)]$

$$f(x^{k+1}) \overset{①}{\le} f(x^k) - \frac{1}{2L} \| \nabla f(x^k) \|^2$$

$$\overset{②}{\le} f(x^k) - \frac{1}{\cancel{2}L} \cancel{2}m [f(x^k) - f(x^*)]$$

$$f(x^{k+1}) \le \left(1 - \frac{m}{L}\right) f(x^k) + \frac{m}{L} f(x^*) = \left(1 - \frac{m}{L}\right)(f(x^k) - f(x^*))$$

$-f(x^*)$ $\qquad$ $-f(x^*)$

$$f(x^{k+1}) - f(x^*) \le \left(1 - \frac{m}{L}\right)\left(f(x^k) - f(x^*)\right)$$

$$\le \cdots$$

$$\le \underbrace{\left(1 - \frac{m}{L}\right)^k}_{\to 0} \left[f(x^0) - f(x^*)\right]$$

$$f(x^k) \longrightarrow f(x^*)$$

$$1 - \frac{m}{L} \in (0,1)$$

$$\frac{m}{L} \approx 0.5 \qquad 1 - \frac{m}{L} \approx 0.5$$

$$\frac{m}{L} = 0.0000001 \qquad 1 - \frac{m}{L} \approx 1$$

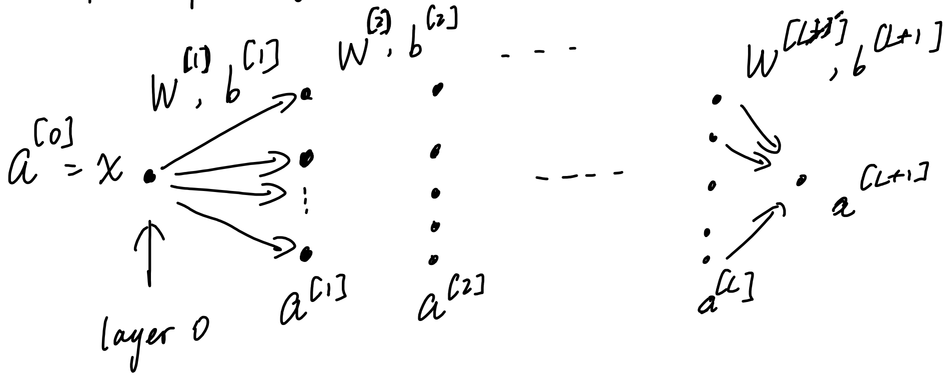
$$k >> 1$$

$$mI \le D^2 f(x) \le LI \qquad \kappa(D^2 f) = \frac{L}{m}$$

"ill conditioned"

Backpropagation. ("Neural Networks")

$a^{[0]} = x$



$$\begin{cases} z^{[l]} = W^{[l]} a^{[l-1]} + b^{[l]} \\ \qquad l = 1, 2, \ldots, L+1 \\ a^{[l]} = \sigma(z^{[l]}) \end{cases}$$

$l = 1, 2, \ldots, L+1$

Loss function

$$C(W, b) = \frac{1}{2}\left(y - a^{[L+1]}\right)^2$$

$W = (W^{[1]} \ldots W^{[L+1]})$ $\quad b = (b^{[1]} \ldots b^{[L+1]})$

$a^{[L+1]} = g((x, y); W, b)$

GD on NN $\left(W^{(k+1)}, b^{(k+1)}\right) = \left(W^{(k)}, b^{(k)}\right) - \alpha \nabla C\left(\underline{W}^{(k)}, \underline{b}^{(k)}\right)$

$$C = C(W, b) = \frac{1}{2}\left(y - a^{[L+1]}\right)^2$$

$$\frac{\partial C}{\partial W_{ij}^{[l]}} \qquad \frac{\partial C}{\partial b_i^{[l]}}$$

Monday, July 27, 2020 9:45 AM

"Backpropagation"

$$\begin{cases} z^{[l+1]} = W^{[l]} a^{[l]} + b^{[l]} \\ a^{[l+1]} = \sigma(z^{[l+1]}) \end{cases} \Bigg\} \; z^{[l+1]} = W^{[l]} \sigma(z^{[l]}) + b^{[l]}$$

$$C = \frac{1}{2}(a^{[L+1]} - y)^2$$

$$= f_1 \circ f_2 \circ \cdots \circ f_n$$

$$a^{[L+1]} = a^{[L+1]}(W, b)$$

$$= \sigma(W^{[l]} a^{[l]} + b^{[l]})$$

$$= \sigma(W^{[l]} (\sigma(W^{[l-1]} a^{[l-1]} + b^{[l-1]})) + b^{[l]})$$

- - -

"error" $\boxed{\delta^{[l]} = \frac{\partial C}{\partial z^{[l]}}}$

$$\delta^{[L+1]}$$

$$\frac{\partial C}{\partial a^{[L+1]}} = a^{[L+1]} - y$$

$$\text{''} \; \frac{\partial C}{\partial z^{[L+1]}} = \frac{\partial C}{\partial a^{[L+1]}} \frac{\partial a^{[L+1]}}{\partial z^{[L+1]}} = (a^{[L+1]} - y) \cdot \sigma'(z^{[L+1]})$$

Monday, July 27, 2020 9:50 AM

$$\frac{\partial C}{\partial W^{[L]}} = \frac{\partial C}{\partial z^{[L+1]}} \frac{\partial z^{[L+1]}}{\partial W^{[L]}} = (a^{[L+1]} - y)\,\sigma'(z^{[L+1]})\, a^{[L]}$$

$$\frac{\partial C}{\partial b^{[L]}} = \frac{\partial C}{\partial z^{[L+1]}} \frac{\partial z^{[L+1]}}{\partial b^{[L]}} = (a^{[L+1]} - y)\,\sigma'(z^{[L+1]})$$

$$\frac{\partial C}{\partial z^{[l]}} = \frac{\partial C}{\partial z^{[l+1]}} \frac{\partial z^{[l+1]}}{\partial z^{[l]}} = \frac{\partial C}{\partial z^{[l+1]}} \frac{\partial z^{[l+1]}}{\partial a^{[l]}} \frac{\partial a^{[l]}}{\partial z^{[l]}}$$

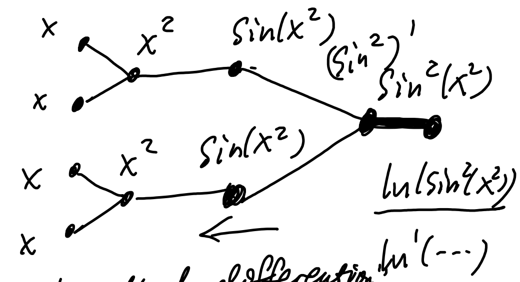
$$\boxed{\delta^{[l]} = \delta^{[l+1]} W^{[l]} \sigma'(z^{[l]})}$$ "Backpropagation"

Monday, July 27, 2020 9:56 AM

$$\delta^{[L+1]} \longrightarrow \delta^{[L]} \longrightarrow \delta^{[L-1]} \longrightarrow \cdots \longrightarrow \delta^{[l]}$$

$$\frac{\partial C}{\partial W^{[l]}} = \frac{\partial C}{\partial z^{[l+1]}} \frac{\partial z^{[l+1]}}{\partial W^{[l]}} = \delta^{[l+1]} a^{[l]}$$

$$\frac{\partial C}{\partial b^{[l]}} = \frac{\partial C}{\partial z^{[l+1]}} \frac{\partial z^{[l+1]}}{\partial b^{[l]}} = \delta^{[l+1]}$$



"computational graph"

$$f(x) = \ln(\sin^2(x^2))$$

$$f'(x) = \ln'(\sin^2)'\sin' \cdots$$

"automatical differentiation"

Accelerated Gradient Descent

f is m-strongly convex $\quad \nabla f$ is L-Lipschitz $\iff mI \le \nabla^2 f \le LI$

$Q \geq 0 \qquad f(x) = \frac{1}{2} x^T Q x - b^T x + c$

$\nabla^2 f(x) = Q$

$\lambda_{min}(Q) = m$

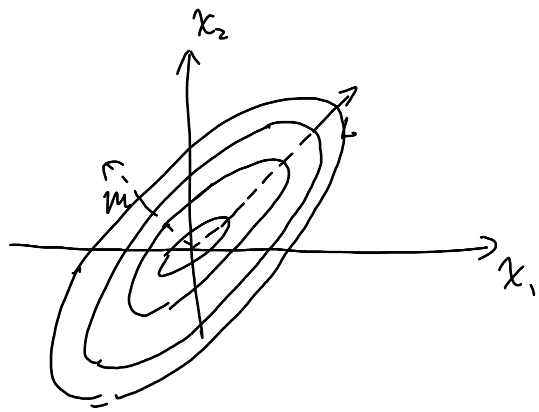
$\lambda_{max}(Q) = L$

$x = (x_1, x_2)^T$

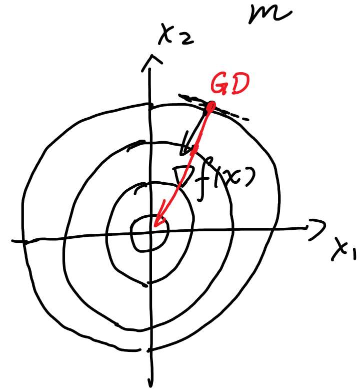
Wednesday, July 29, 2020 8:44 AM

$$f(x) = \frac{1}{2} x^T Q x = \frac{1}{2}\left[q_{11} x_1^2 + q_{22} x_2^2 + 2\, q_{12} x_1 x_2 \right]$$

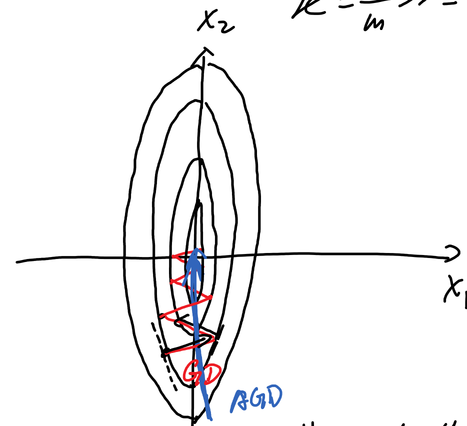
"level set"



$$\kappa = \frac{L}{m}$$



$$\kappa = \frac{L}{m} \gg 1$$



"Markov"

$$x^{k+1} = x^k - \alpha \boxed{\nabla f(x^k)}$$

"inertia" $\Longleftrightarrow$ "acceleration" $\quad F = ma = m\dfrac{d^2x}{dt^2}$

<u>GD</u>

x^0

$$x^{k+1} = x^k - \alpha \nabla f(x^k)$$

$$\frac{x^{k+1} - x^k}{\Delta t} = -\cancel{\alpha} \nabla f(x^k)$$

$\cancel{\Delta t}$

$\updownarrow$

$$\frac{dx}{dt} = -\nabla f(x)$$

Gradient Flow

<u>AGD</u> $\quad \mu$ — Heavy Ball

$$\mu \frac{d^2x}{dt^2} = -\nabla f(x) - \mu b \frac{dx}{dt}$$

$$\mu \frac{x(t+\Delta t) - 2x(t) + x(t-\Delta t)}{(\Delta t)^2}$$

$$= -\nabla f(x) - \mu b \frac{x(t+\Delta t) - x(t)}{\Delta t}$$

$$\mu \frac{x(t+\Delta t)-2x(t)+x(t-\Delta t)}{(\Delta t)^2} = -\nabla f(x^{(t)}) - \mu b \frac{x(t+\Delta t)-x(t)}{\Delta t}$$

$$\rightsquigarrow \quad x(t+\Delta t) = \text{----}\; x(t) + \text{---}\; x(t-\Delta t) + \cdots \nabla f \qquad (\star)$$

$$\mu\left[\frac{1}{(\Delta t)^2}+\frac{b}{\Delta t}\right]x(t+\Delta t) = -\nabla f(x) + \left[\frac{2\mu}{(\Delta t)^2}+\frac{\mu b}{\Delta t}\right]x(t) - \frac{\mu}{(\Delta t)^2}x(t-\Delta t)$$

$$(1+b\Delta t)\mu\, x(t+\Delta t) = -(\Delta t)^2\nabla f(x) + \mu[2+b\Delta t]\,x(t) - \mu\, x(t-\Delta t)$$

$$x(t+\Delta t) = -\overbrace{\frac{(\Delta t)^2}{(1+b\Delta t)\mu}}^{=\alpha}\nabla f(x(t)) + \frac{1+b\Delta t+1}{1+b\Delta t}x(t) - \overbrace{\frac{1}{1+b\Delta t}}^{=\beta}x(t-\Delta t)$$

$$x(t+\Delta t) = -\alpha\nabla f(x(t)) + x(t) + \beta(x(t)-x(t-\Delta t)) \qquad (AGD)$$

$$x(t+\Delta t) = -\alpha\nabla f(x(t)) + x(t) \qquad (GD)$$

Polyak's Heavy Ball method (GD with momentum)

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta \underbrace{(x^k - x^{k-1})}_{\text{momentum}} , \quad x^0 = x^{-1}$$

α: learning rate

β: momentum constant.

$x \quad m\frac{dx}{dt}$

"momentum" $p^k = x^{k+1} - x^k$

$$\begin{cases} p^k = -\alpha \nabla f(x^k) + \beta p^{k-1} \\ x^{k+1} = x^k + p^k \end{cases}$$

$$\left.\begin{array}{l} \underline{GD} \\ \underline{AGD} \end{array}\right\} \; x^{k+1} = x^k + p^k \; \left\} \begin{cases} GD \to p^k = -\alpha \nabla f(x^k) \\ AGD \to p^k = -\alpha \nabla f(x^k) + \beta p^{k-1} \end{cases}\right.$$

$$\begin{aligned} p^k &= -\alpha \nabla f(x^k) - \beta\alpha \nabla f(x^{k-1}) + \beta^2 p^{k-2} \\ &= -\alpha \nabla f(x^k) - \beta\alpha \nabla f(x^{k-1}) - \beta^2\alpha \nabla f(x^{k-2}) + \beta^3 p^{k-3} \\ &= \cdots \\ &= -\alpha \nabla f(x^k) - \beta\alpha \nabla f(x^{k-1}) - \beta^2\alpha \nabla f(x^{k-2}) - \cdots \end{aligned}$$

$\beta^m \alpha \nabla f(x^{k-m})$

"moving-average"

Polyak's Heavy Ball

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1})$$

"Explicit Scheme"

$$x^{k+1} = F(x^k, x^{k-1}, \dots)$$

$$x^{k+1} = x^k - \alpha \nabla f(x^k + \beta(x^k - x^{k-1})) + \beta(x^k - x^{k-1})$$

Nesterov's Accelerated GD

"Implicit Scheme"

$$x^{k+1} = F(x^{k+1}, x^k, x^{k-1} \dots)$$

"

$$x^{k+1} = x^k - \alpha \nabla f(x^{k+1}) + \beta(x^k - x^{k-1})$$

"

$$x^k - \alpha \nabla f(x^k + \beta(x^k - x^{k-1}) \cancel{- \alpha \nabla f(x^{k+1})}) + \beta(x^k - x^{k-1})$$

Polyak $x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1})$

Nesterov $x^{k+1} = x^k - \underline{\alpha} \nabla f(x^k + \underline{\beta}(x^k - x^{k-1})) + \beta(x^k - x^{k-1})$

Meta-Theorem α, β tuned appropriately

f is m-strongly convex ∇f is L-Lipschitz

$$f(x^{k+1}) - f(x^*) \le \left(1 - \sqrt{\frac{m}{L}}\right)^k \left[f(x^0) - f(x^*) + \frac{m}{2}\|x^0 - x^*\|^2 \right]$$

	$\kappa = \frac{L}{m}$
GD	$1 - \frac{1}{\kappa}$
AGD	$1 - \frac{1}{\sqrt{\kappa}}$

$$\kappa = 10^8$$

$$1 - 10^{-8} \approx 1$$

$$1 - 10^{-4} \ll 1 - 10^{-8}$$

Gradient - Based Methods

GD $\quad x^{k+1} = x^k - \alpha \nabla f(x^k)$

AGD $\quad x^{k+1} = x^k - \alpha \nabla f(x^k + \beta(x^k - x^{k-1})) + \beta(x^k - x^{k-1})$

ERM $\quad (a_i, b_i)$

input label

$l_i(x) = l(g(a_i; x), b_i) \overset{e.g.}{=} |g(a_i; x) - b_i|^2$

$L_n(x) = \frac{1}{n} \sum_{i=1}^{n} l_i(x)$

Empirical Risk

$x_n^* = \operatorname{argmin}_x L_n(x)$

Gradient-Based Optimization applied to ERM

$$\nabla L_n(x) = \frac{1}{n} \sum_{i=1}^{n} \nabla l_i(x)$$

GD $$x^{k+1} = x^k - \frac{\alpha}{n} \sum_{i=1}^{n} \nabla l_i(x^k) \qquad k = 1 \dots T \gg 1$$

Large-Scale ML problems $x \in \mathbb{R}^D$ $D \gg 1$

training data (a_i, b_i) $i = 1 \dots n$ $n \gg 1$

IFO (Incremental First-Order Oracle)

IFO[GD] $\longrightarrow$ $n \times T$

Stochastic Gradient Descent (SGD)

<u>GD</u> $$x^{k+1} = x^k - \frac{\alpha}{n}\sum_{i=1}^{n} \nabla l_i(x^k) \qquad B = n$$

<u>minibatch-SGD</u> minibatch $\boxed{1 \le B \le n}$ take randomly uniformly size B subsets from $\{1 \dots n\}$

$\updownarrow$

full-batch $\{1, 2, 3, \dots, n-1, n\}$

e.g. $n = 3$ $\{1, 2, 3\}$

$B = 2$ $\underline{\{1, 2\}}$ $\underline{\{1, 3\}}$ $\underline{\{2, 3\}}$

$\uparrow$ $p = 1/3$ $\uparrow$ $p = 1/3$ $\uparrow$ $p = 1/3$

$\underline{B = 1}$ $\{1 \dots n\}$

mini-batch i taken randomly uniformly from $\{1 \dots n\}$

$\mathcal{B} \subset \{1 \dots n\}$ s.t. $|\mathcal{B}| = B$ and $\mathcal{B}$ taken randomly uniformly

$\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_K, \dots$ i.i.d

$B \underline{<< n}$

mini-batch SGD

$\underline{\text{IFO}}[SGD] = B \times T$

$g_B(x) = \frac{1}{B} \sum_{i \in B} \nabla l_i(x)$ "gradient estimator"

$$x^{k+1} = x^k - \frac{\alpha}{B} \sum_{i \in B_k} \nabla l_i(x^k) = x^k - \alpha g_{B_k}(x^k)$$

$$L_n(x) = \frac{1}{n} \sum_{i=1}^{n} l_i(x)$$

$\underline{\underline{B = 1}}$ $\quad x^{k+1} = x^k - \alpha \underline{\underline{\nabla l_{i_k}}(x^k)}$

$i_1, i_2, \ldots, i_k, \ldots$ is drawn randomly uniformly from $\{1 \ldots n\}$ in an i.i.d way

$\mathbb{P}(i = k) = \frac{1}{n}$
$k = 1 \ldots n$

$$L_n(x) = \mathbb{E}\, l_i(x)$$

SGD

GD

Monday, August 3, 2020 8:53 AM

$n=3$

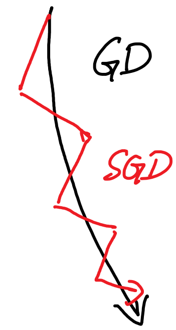
$$x^1 = x^0 - \alpha \nabla \ell_2(x^0)$$
$$x^2 = x^1 - \alpha \nabla \ell_1(x^1)$$
$$x^3 = x^2 - \alpha \nabla \ell_3(x^2)$$
$$\vdots$$

$$x^1 = x^0 - \alpha \frac{\nabla \ell_1(x^0) + \nabla \ell_2(x^0) + \nabla \ell_3(x^0)}{3}$$
$$x^2 = x^1 - \alpha \frac{\nabla \ell_1(x^1) + \nabla \ell_2(x^1) + \nabla \ell_3(x^1)}{3}$$
$$x^3 = x^2 - \alpha \frac{\nabla \ell_1(x^2) + \nabla \ell_2(x^2) + \nabla \ell_3(x^2)}{3}$$
$$\vdots$$

In general if $\mathcal{B}$ is taken randomly uniformly from size B subsets of $\{1, \ldots, n\}$ $\quad \binom{n}{B} = \frac{n!}{B!(n-B)!}$

then $$\frac{1}{n}\sum_{i=1}^{n} \nabla \ell_i(x) = \mathbb{E}_{\mathcal{B}}\left(\frac{1}{B}\sum_{i\in\mathcal{B}} \nabla \ell_i(x)\right)$$

"unbiasedness" $$\nabla L_n(x) = \mathbb{E}_{\mathcal{B}}\left(g_{\mathcal{B}}(x)\right) \qquad g_{\mathcal{B}} = \frac{1}{B}\sum_{i\in\mathcal{B}} \nabla \ell_i(x)$$



GD $$x^{k+1} = x^k - \alpha \nabla L_n(x^k)$$

SGD $$x^{k+1} = x^k - \alpha\, g_{B_k}(x^k)$$

$$\mathbb{E}\, g_{B_k}(x^k) = \nabla L_n(x^k)$$

$$\mathbb{E}\left[g_{B_k}(x^k) - \mathbb{E}\, g_{B_k}(x^k)\right]^2 \uparrow \text{ as } B \downarrow$$

"Convergence"

Monday, August 3, 2020 9:13 AM

$$\mathbb{E}\, g_{\zeta_k}(x^k) = \nabla f(x^k)$$

"unbiased"

SGD $\quad x^{k+1} = x^k - \alpha \underbrace{g_{\zeta_k}(x^k)}_{\hookrightarrow \text{ Gradient Estimator}}$

$\zeta_1 \ldots \zeta_k, \ldots$ is i.i.d $\qquad g_{\zeta_k}$

$x^* \qquad \mathbb{E} = \mathbb{E}^k$

f is m-strongly convex

$$\|x^{k+1} - x^*\|^2 = \|x^k - \alpha g_{\zeta_k}(x^k) - x^*\|^2$$

$$= \|x^k - x^*\|^2 - 2\alpha \langle x^k - x^*, g_{\zeta_k}(x^k)\rangle + \alpha^2 \|g_{\zeta_k}(x^k)\|^2$$

$$\mathbb{E}\left(\|x^{k+1} - x^*\|^2 \,\middle|\, x^k\right) = \|x^k - x^*\|^2 - 2\alpha \langle x^k - x^*, \nabla f(x^k)\rangle + \alpha^2 \mathbb{E}_{\zeta}\left(\|g_{\zeta_k}(x^k)\|^2\right)$$

$$\langle Df(x),\ x-x^* \rangle \geq m\|x-x^*\|^2 \qquad (\text{Ex.})$$

Hint $D^2 f \geq mI$.

$$\mathbb{E}\left[\|x^{k+1}-x^*\|^2 \,\middle|\, x^k\right] \leq \|x^k-x^*\|^2 - 2\alpha m\|x^k-x^*\|^2 + \alpha^2 \underline{\mathbb{E}_\xi \|g_\xi(x^k)\|^2}$$

$$\mathbb{E}\left[\mathbb{E}[A \mid B]\right] = \mathbb{E}A \qquad (\text{“telescoping”})$$

$$\mathbb{E}\|x^{k+1}-x^*\|^2 \leq (1-2\alpha m)^k\, \mathbb{E}\|x^0-x^*\|^2 + 2\alpha^2 \underbrace{\left[\mathrm{Var}(g_\xi)+\mathbb{E}\|Df\|^2\right]}$$

$$\mathbb{E}g_\xi = Df \qquad \mathbb{E}\|g_\xi\|^2 = \mathbb{E}\|(g_\xi - Df)+Df\|^2 \leq 2\,\underline{\mathrm{Var}(g_\xi)} + 2\mathbb{E}\|Df\|^2$$

$$\mathbb{E}\|x^{k+1}-x^*\|^2 \le (1-2\alpha m)\,\mathbb{E}\|x^k-x^*\|^2 + B^2$$

$$\le (1-2\alpha m)^2\,\mathbb{E}\|x^{k-1}-x^*\|^2 + (1-2\alpha m)B^2 + B^2$$

$$\le \cdots$$

$$\le (1-2\alpha m)^k\,\underline{\mathbb{E}\|x^1-x^*\|^2} + \underbrace{(1-2\alpha m)^{k-1}B^2 + \cdots + (1-2\alpha m)B^2 + B^2}$$

$$\mathbb{E}\|x^k-x^*\|^2 \le \lambda^k \cdot C_0 + C_1$$

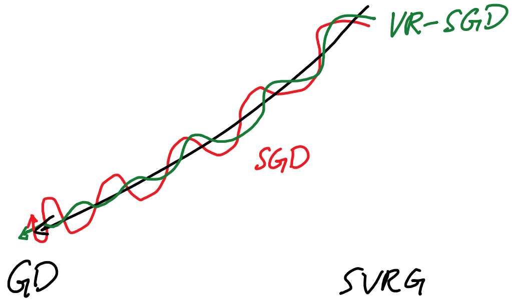
$0<\lambda<1$

$C_1^k \to 0$ as $k\to\infty$

(VR)

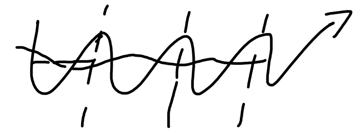
$$\frac{B^2}{2\alpha m} = C_1 \quad \text{(} \nparallel \text{)}$$

Monday, August 3, 2020 9:30 AM



SVRG

- Variance-Reduced SGD   2014?   Johnson-Zhang



SGD $x^{k+1} = x^k - \alpha \nabla \ell_{i_k}(x^k)$

VR-SGD   "Snapshots" $x^0 \to x^1 \to \cdots \to x^M \rightsquigarrow \tilde{x}^{(1)}$

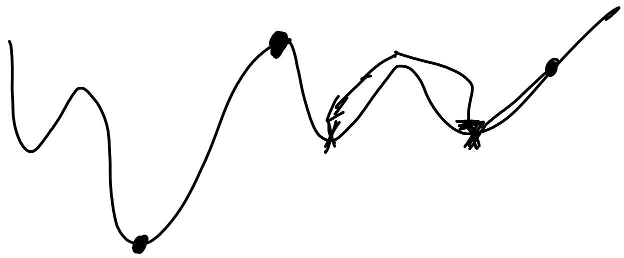
$\to x^{M+1} \to x^{M+2} \to \cdots \to x^{2M} \rightsquigarrow \tilde{x}^{(2)}$

$$x^{k+1} = x^k - \alpha \Big(\underbrace{\nabla \ell_{i_k}(x^k) - \nabla \ell_{i_k}(\tilde{x}) + \frac{1}{n}\sum_{i=1}^{n} \nabla \ell_i(\tilde{x})}_{\text{VR gradient estimator}} \Big)$$

- Non-Convex Optimization

"Implicit Regularization"

SGD



saddle-pt

"escape from saddles"

Practical Issues

Adaptive Methods

GD

gradient

Momentum

Stochastic Gradients

$\nabla \longrightarrow \hat{\nabla}$