EPFL CS-439 (optim. for ML) spring 2022
template from https://github.com/dcetin/eth-cs-notes/ (pai-cheatsheet)

GD basics $\mathbf{g}_t := \nabla f(\mathbf{x}_t)$

$$\sum_{t=0}^{T-1} \mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) = \frac{\gamma}{2}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\left(\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 - \|\mathbf{x}_T - \mathbf{x}^\star\|^2\right)$$

Lipschitz convex: 1/eps^2

$$\gamma := \frac{R}{B\sqrt{T}} \qquad \frac{1}{T}\sum_{t=0}^{T-1}\left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) \le \frac{RB}{\sqrt{T}}$$

Smooth convex: 1/eps

$$\gamma := \frac{1}{L} \qquad f(\mathbf{x}_T) - f(\mathbf{x}^\star) \le \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$$

suff. descent: $$f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2$$

Smooth convex accelerated: 1/sqrt(eps)

Nesterov's accelerated gradient descent ('83)

Smooth strongly-convex: (L/mu) log(1/eps)

$$\gamma := \frac{1}{L} \qquad \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \le \left(1 - \frac{\mu}{L}\right)\|\mathbf{x}_t - \mathbf{x}^\star\|^2$$

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \le \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$$

PGD $\mathbf{g}_t := \nabla f(\mathbf{x}_t)$ $\mathbf{x}_{t+1} := \Pi_X(\mathbf{y}_{t+1})$

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 + \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2 \le \|\mathbf{y}_{t+1} - \mathbf{x}^\star\|^2$$

(second term can be seen as noise, often cancels out)

Lipschitz convex: idem, 1/eps^2

Technically, only need bounded gradient ⊋ lipschitz (X closed)

Smooth convex: 1/eps

suff. "descent" $f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2$

Smooth strongly-convex: (L/mu) log(1/eps)

square distance to OPT still geom. decreasing, but

$$\begin{aligned} f(\mathbf{x}_T) - f(\mathbf{x}^\star) \ &\le\ \|\nabla f(\mathbf{x}^\star)\|\left(1 - \frac{\mu}{L}\right)^{T/2}\|\mathbf{x}_0 - \mathbf{x}^\star\| \\ &+ \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0. \end{aligned}$$

Proximal grad: f=g+h, $\mathbf{x}_{t+1} := \mathrm{prox}_{h,\gamma}(\mathbf{x}_t - \gamma\nabla g(\mathbf{x}_t))$

$$\mathrm{prox}_{h,\gamma}(\mathbf{z}) := \underset{\mathbf{y}}{\mathrm{argmin}}\left\{\frac{1}{2\gamma}\|\mathbf{y} - \mathbf{z}\|^2 + h(\mathbf{y})\right\}$$ non-expansive

g smooth, g,h convex, $\gamma := \frac{1}{L}$: $f(\mathbf{x}_T) - f(\mathbf{x}^\star) \le \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$

Subgradient descent $\mathbf{g}_t \in \partial f(\mathbf{x}_t)$

Lipschitz convex: idem, 1/eps^2

Tame strongly-convex $B = \max_{t=1}^T \|\mathbf{g}_t\|$: 1/eps

$$\gamma_t := \frac{2}{\mu(t+1)} \qquad f\left(\frac{2}{T(T+1)}\sum_{t=1}^T t\cdot\mathbf{x}_t\right) - f(\mathbf{x}^\star) \le \frac{2B^2}{\mu(T+1)}$$

Reason for step-size choice: must multiply by t before telescoping

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \le \frac{B^2\gamma_t}{2} + \frac{(\gamma_t^{-1} - \mu)}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \frac{\gamma_t^{-1}}{2}\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2$$

Lower bound (Nesterov): ∃ f B-lipschitz s.t

for any subgradient method, $$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \ge \frac{RB}{2(1+\sqrt{T+1})}$$

SGD

Smoothness never helps! Only bounded gradients ("tame")

(Smoothness may help when doing variance-reduced SGD)

Lipschitz convex: idem, 1/eps^2

Technically, only need bounded stoch. gradients $\mathbb{E}\left[\|\mathbf{g}_t\|^2\right] \le B^2$

Tame strongly-convex $B = \max_{t=1}^T \mathbb{E}\left[\|\mathbf{g}_t\|\right]$: 1/eps

$$\gamma_t := \frac{2}{\mu(t+1)} \qquad \mathbb{E}\left[f\left(\frac{2}{T(T+1)}\sum_{t=1}^T t\cdot\mathbf{x}_t\right) - f(\mathbf{x}^\star)\right] \le \frac{2B^2}{\mu(T+1)}$$

(same proof as in subgradient descent with expectations)

Mini-batch reduces variance: $\mathbb{E}\left[\left\|\tilde{\mathbf{g}}_t - \nabla f(\mathbf{x}_t)\right\|^2\right] \le \frac{B^2}{m}$

where batch size=m and $\mathbb{E}\left[\|\mathbf{g}_t\|^2\right] \le B^2$

Nonconvex functions

Def. of smooth is still only an upper bound!

TL;DR: $\|\nabla f(\mathbf{x}_t)\|^2$ →0 at same rate as $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ for convex

Smooth: 1/eps ON AVERAGE

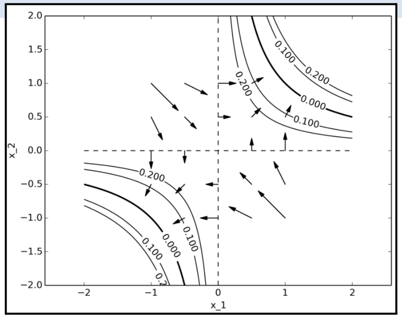
$$\gamma := \frac{1}{L} \qquad \frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 \le \frac{2L}{T}\left(f(\mathbf{x}_0) - f(\mathbf{x}^\star)\right)$$

No-overshoot ppty: ∄ critical pt on segment [xt,xt+1]

that $\nabla f(\mathbf{x}) \neq \mathbf{0}$, i.e. $\mathbf{x}$ is not a critical point. Suppose that f is smooth with parameter L over the line segment connecting $\mathbf{x}$ and $\mathbf{x}' = \mathbf{x} - \gamma\nabla f(\mathbf{x})$, where $\gamma = 1/L' < 1/L$. Then $\mathbf{x}'$ is also not a critical point.

The example

$$f(\mathbf{x}) := \frac{1}{2}\left(\prod_{k=1}^d x_k - 1\right)^2$$



Not globally smooth, yet

$$f(\mathbf{x}_T) \le \left(1 - \frac{\delta^2}{3c^4}\right)^T f(\mathbf{x}_0).$$

Newton's method $\mathbf{x}_{t+1} := \mathbf{x}_t - \nabla^2 f(\mathbf{x}_t)^{-1}\nabla f(\mathbf{x}_t)$

$$\mathbf{x}_{t+1} = \underset{\mathbf{x}\in\mathbb{R}^d}{\mathrm{argmin}}\ f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top(\mathbf{x} - \mathbf{x}_t) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_t)^\top\nabla^2 f(\mathbf{x}_t)(\mathbf{x} - \mathbf{x}_t)$$

(No step-size)

Rk. Newton's method is **affine-invariant**.

Thm. Suppose ∃ ball around x* where (for spectral norm)

$$\|\nabla^2 f(\mathbf{x})^{-1}\| \le \frac{1}{\mu} \text{ and } \|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\| \le B\|\mathbf{x} - \mathbf{y}\|$$

Then if x0∈ball, $$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\| \le \frac{B}{2\mu}\|\mathbf{x}_t - \mathbf{x}^\star\|^2.$$

Corr. If x0∈ball and $\|\mathbf{x}_0 - \mathbf{x}^\star\| \le \frac{\mu}{B}$, $\|\mathbf{x}_T - \mathbf{x}^\star\| \le \frac{\mu}{B}\left(\frac{1}{2}\right)^{2^T-1}$

i.e to get $\|\mathbf{x}_T - \mathbf{x}^\star\| < \varepsilon$, only need $T = \log\log(\frac{1}{\varepsilon})$

Local quadratic convergence ("double the number of correct digits in each iteration")

- affine invariant
- converge in 1 step for quadratics

Quasi-Newton $\mathbf{x}_{t+1} = \mathbf{x}_t - H_t^{-1}\nabla f(\mathbf{x}_t)$,

with H symmetric s.t secant condition:

$$\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}) = H_t(\mathbf{x}_t - \mathbf{x}_{t-1})$$

In 1D, only one secant method $$x_{t+1} := x_t - f'(x_t)\frac{x_t - x_{t-1}}{f'(x_t) - f'(x_{t-1})}$$

Greenstadt family, of which (L-)BFGS

Newton ∈ Quasi-Newton ⇔ f nondegen. quadratic

Coordinate descent $\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_i \nabla_i f(\mathbf{x}_t)\,\mathbf{e}_i$

PL inequality: $\frac{1}{2}\|\nabla f(\mathbf{x})\|^2 \ge \mu(f(\mathbf{x}) - f(\mathbf{x}^\star))$

μ-strongly-convex ⇒ μ-PL

E.g f(x)=x1^2 is 1-PL but not SC

E.g $f(\mathbf{x}) := g(A\mathbf{x})$ for strongly convex g and arbitrary matrix A

GD on smooth + PL: (L/mu) log(1/eps)

(Exact same proof as for smooth+SC)

Randomized CD $i \in [d]$ uniformly

If f is (L, …, L)-coord-wise-smooth and μ-PL, $\gamma_i = \frac{1}{L}$,

$$\mathbb{E}[f(\mathbf{x}_T) - f(\mathbf{x}^\star)] \le \left(1 - \frac{\mu}{dL}\right)^T (f(\mathbf{x}_0) - f(\mathbf{x}^\star))$$

Importance sampling CD $i \in [d]$ with probability $\frac{L_i}{\sum_{j=1}^d L_j}$

If f is (L1, …, Ld)-coord-wise-smooth and μ-PL, stepsize $\frac{1}{L_i}$,

$$\mathbb{E}[f(\mathbf{x}_T) - f(\mathbf{x}^\star)] \le \left(1 - \frac{\mu}{d\bar{L}}\right)^T (f(\mathbf{x}_0) - f(\mathbf{x}^\star))\ ; \qquad \bar{L} = \frac{1}{d}\sum_{i=1}^d L_i$$

E.g f(x)=x1^2 is (2, 0, …, 0)-smooth so L=2 and ¯L=2/d

Steepest CD aka Gauss-Southwell $i = \operatorname{argmax}_{i\in[d]} |\nabla_i f(\mathbf{x}_t)|$

If f is (L, …, L)-coord-wise-smooth and μ-PL, same bound as for Randomized CD

→ strictly worse bound, as per-iteration cost is ~d

If f is (L, …, L)-coord-wise-smooth and μ1-PL **w.r.t l1 norm**,

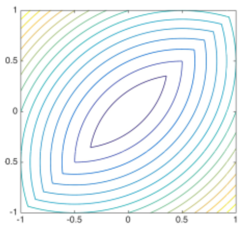
$\gamma_i = \frac{1}{L}$, $f(\mathbf{x}_T) - f(\mathbf{x}^\star) \le \left(1 - \frac{\mu_1}{L}\right)^T (f(\mathbf{x}_0) - f(\mathbf{x}^\star))$

Rk: μ1-SC w.r.t l1 norm ⇒ μ1-PL w.r.t l1 norm

$$\tfrac{1}{2}\|\nabla f(\mathbf{x})\|_\infty^2 \ge \mu_1(f(\mathbf{x}) - f(\mathbf{x}^\star))$$

Greedy CD: line-search $\mathbf{x}_{t+1} := \operatorname{argmin}_{\lambda\in\mathbb{R}} f(\mathbf{x}_t + \lambda\mathbf{e}_i)$

May fail: $\|\mathbf{x}\|^2 + |x_1 - x_2|$



Thm. If f = g+h, g convex diffble and $h(\mathbf{x}) = \sum_i h_i(x_i)$ with hi convex then xt+1=xt ⇒ xt global min of f

Frank-Wolfe aka conditional gradient

$$\mathbf{s} := \mathrm{LMO}_X(\nabla f(\mathbf{x}_t))$$
$$\mathbf{x}_{t+1} := (1-\gamma_t)\mathbf{x}_t + \gamma_t\mathbf{s}$$

Lin. Min. Oracle $\mathrm{LMO}_X(\mathbf{g}) := \operatorname{argmin}_{\mathbf{z}\in X} \mathbf{g}^\top\mathbf{z}$

If X = conv(A), then LMO_X(g) ∈ A ("atoms")

Examples	$\mathcal{A}$	$\lvert\mathcal{A}\rvert$	dim.	$\mathrm{LMO}_X(\mathbf{g})$
L1-ball	$\{\pm\mathbf{e}_i\}$	$2d$	d	$\pm\mathbf{e}_i$ with $\operatorname{argmax}_i \lvert g_i\rvert$
Simplex	$\{\mathbf{e}_i\}$	d	d	$\mathbf{e}_i$ with $\operatorname{argmin}_i g_i$
Spectahedron	$\{\mathbf{x}\mathbf{x}^\top, \|\mathbf{x}\| = 1\}$	∞	d^2	$\operatorname{argmin}_{\|\mathbf{x}\|=1} \mathbf{x}^\top G\mathbf{x}$
Norms	$\{\mathbf{x}, \|\mathbf{x}\| \le 1\}$	∞	d	$\operatorname{argmin}_{\mathbf{s},\|\mathbf{s}\|\le 1} \langle\mathbf{s},\mathbf{g}\rangle$
Nuclear norm	$\{Y, \|Y\|_* \le 1\}$	∞	d^2	..

(Spectrahedron: PSD matrices with trace=1;

$\mathrm{LMO}_X(G) = \mathbf{s}_1\mathbf{s}_1^\top$ via eigenvector)

Duality gap $g(\mathbf{x}) := \nabla f(\mathbf{x})^\top(\mathbf{x} - \mathbf{s})$

g(x) ≥ f(x)-f(x*) and g(x*) = 0

Curvature constant

$$C_{(f,X)} := \sup_{\substack{\mathbf{x},\mathbf{s}\in X,\gamma\in(0,1]\\ \mathbf{y}=(1-\gamma)\mathbf{x}+\gamma\mathbf{s}}} \frac{1}{\gamma^2}\left(f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x})\right)$$

If f is L-smooth, then $C_{(f,X)} \le \frac{L}{2}\operatorname{diam}(X)^2$.

Allows to capture that Frank-Wolfe algo is **affine-invariant**.

Smooth convex: 1/eps

Analysis is quite different from (S/P/prox)GD, CD, Newton.

Thm. If X convex compact, f convex, C_(f,X) < ∞,

step-size $\gamma_t = 2/(t+2)$ (indep of params!)

or $\operatorname{argmin}_{\gamma\in[0,1]} f((1-\gamma)\mathbf{x}_t + \gamma\mathbf{s})$ or $\min\left(\frac{g(\mathbf{x}_t)}{L\,\|\mathbf{s} - \mathbf{x}_t\|^2}, 1\right)$,

then $f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top\gamma_t(\mathbf{s} - \mathbf{x}) + \gamma_t^2 C_{(f,X)}$

and so $h(\mathbf{x}_{t+1}) \le (1-\gamma_t)h(\mathbf{x}_t) + \gamma_t^2 C$

and so $f(\mathbf{x}_T) - f(\mathbf{x}^\star) \le \frac{4C_{(f,X)}}{T+1}$.

Thm ("cv" of duality gap). Under the same conditions,

there exists 1≤t≤T s.t $g(\mathbf{x}_t) \le \frac{27/2 \cdot C_{(f,X)}}{T+1}$

Zero-th-order/gradient-free optim

Random search: pick a random direction $\mathbf{d}_t \in \mathbb{R}^d$

$\gamma := \operatorname{argmin}_{\gamma\in\mathbb{R}} f(\mathbf{x}_t + \gamma\mathbf{d}_t)$ (line-search)

$\mathbf{x}_{t+1} := \mathbf{x}_t + \gamma\mathbf{d}_t$

(step-size: no other choice than line-search!)

Convergence rates: same as GD with optimal step-size, with slow-down factor of d

Smooth convex: T < dL/eps

Smooth strongly-convex: T < dL/mu log(1/eps)

Misc

In finite dim, convex => cont. and difble almost everywhere.

If param e.g $\gamma := \frac{1}{L}$ unknown, use doubling trick.

Using line-searched step-size, can only do better than fixed step-size.

To prove SC => PL: min over y in

f(y) > f(x) + g(x)*(y-x)+mu/2 |y-x|2

similarly for non-l2 norms

similarly, can prove L-smooth => "lower-PL"

For any convex L-smooth f,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 \le 2L\left(f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\top(\mathbf{x} - \mathbf{y})\right)$$

(proof: L-smooth => "lower-PL" on tilted h(x)=f(x)-g(y)*x)