Basics

(! General !)
Constrained opt: $\nabla f(x^*) = 0$ not required $\rightarrow$ optimality cond! Check that x^* and iterates are feasible!
Remove const terms in minimization.
For upper bounds: Remove subtractions of non-neg terms & use monotonicity of functions.
split norm sum $\|x \pm y\|^2 = \|x\|^2 + \|y\|^2 \pm 2\langle x, y\rangle$
max $\geq$ avg: $\max_i x_i^2 \geq \frac{1}{d}\|x\|_2^2 = \frac{1}{d}\sum x_i^2$
Ineq. with e.g. indicator func: make case distinction.

(Triangle ineq.) $\|x + y\| \leq \|x\| + \|y\|$
(reverse tri ineq) $|\|x\| - \|y\|| \leq \|x - y\|$

(Parallelogram law)
$\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2$

(Law of cosines)
$\|x - y\|^2 = \|x\|^2 + \|y\|^2 - 2\langle x, y\rangle$

(Cauchy-Schwarz) $|\langle x, y\rangle| \leq \|x\|\,\|y\|$

(Jensen's inequality) f conv, $\sum \lambda_i = 1$, $x_i \in \mathbf{dom}(f)$

$$f\left(\sum \lambda_i x_i\right) \leq \sum \lambda_i f(x_i)$$

$\Rightarrow f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$

(Convex set) $\lambda x + (1-\lambda)y \in C \quad \forall x, y \in C, \lambda \in [0,1]$

(Spectral norm) $\|A\| = \max_{\|x\|=1}\|Ax\|$
Consequence: $\|Ax\| \leq \|A\|\,\|x\|$

(Differentiability) $f : \mathbf{dom}(f) \subseteq \mathbb{R}^d \rightarrow \mathbb{R}^m$ is called diff'able at x in the interior of $\mathbf{dom}(f)$ if $\exists A \in \mathbb{R}^{m\times d}$ and $r : \mathbb{R}^d \rightarrow \mathbb{R}^m$ s.t. $\forall y$ in neighborhood of x:

$$f(y) = f(x) + A(y-x) + r(y-x) \text{with} \lim_{v\rightarrow 0}\frac{\|r(v)\|}{\|v\|} = 0$$

We then define $Df(x)_{ij} = (\partial f_i/\partial x_j)(x)$.

(Lipschitz) f diff'able, $\mathbf{dom}(f)$ convex, $B \in \mathbb{R}_+$. Following is equiv:

$$\|f(x) - f(y)\| \leq B\|x - y\| \;\; (f \text{ is } B\text{-Lipschitz})$$
$$\|Df(x)\| \leq B \text{ (bounded differential)}$$

(Young's inequality) $p, q > 0$ s.t. $1/p + 1/q = 1$ and $a, b \geq 0$

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}, \quad ab \leq \frac{a^2}{2} + \frac{b^2}{2}$$

2nd part $p = q = 2$. Equality holds iff $a^p = b^q$.

Hölder's ineq: $u^\top v \leq \|u\|_\infty \|v\|_1$ **AM-GM:** $n^{-1}\sum x_i \geq \sqrt[n]{\Pi x_i}$ $\Rightarrow$ w/ CS: $|x^\top y| \leq (\|x\|/\sqrt{c})(\|y\|\sqrt{c}) \leq \frac{1}{2}(\|x\|^2/c + c\|y\|^2)$

Norms and seminorms are convex.

Basic inequalities: $\ln(1+x) \leq x; 1 - x \leq e^{-x}$; $\|x\|_2 \leq \|x\|_1 \leq \sqrt{d}\|x\|_2$; $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{d}\|x\|_\infty$

Hypograph: $\mathbf{hyp}f = \{(x,t) \mid f(x) \leq t\}$, epigraph: $\mathbf{epi}f = \{(x,t) \mid f(x) \geq t\}$

Differentiation: $g = Ax + b \Rightarrow \nabla(f \circ g)(x) = A^\top \nabla f(Ax + b)$; $f = x^\top Qx + b^\top x + c \Rightarrow \nabla f(x) = 2Qx + b$; $\nabla x^\top A = A$; $\nabla a^\top x = \nabla x^\top a = a$; $\nabla b^\top Ax = A^\top b$; $\nabla x^\top x = 2x$; $\nabla_w \|y - Xw\|_2^2 = 2X^\top(Xw - y)$

Basic diff: $(fg)' = f'g + fg'$; $(f/g)' = (f'g - fg')/g^2$; $(f \circ g)' = f'(g)g'$

Convex Functions

Convex functions are continuous: $\mathbf{dom}(f)$ open, f convex $\Rightarrow f$ continuous. (proof not obv)

(Convex function) $\forall x, y \in \mathbf{dom}(f)$ conv, $\lambda \in [0,1]$

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$$
$$f(y) \geq f(x) + \nabla f(x)^\top(y - x)$$
$$y^\top \nabla^2 f(x) y \geq 0$$

1oc requires ∇f to exist at every point and $\mathbf{dom}(f)$ open. 1oc is equivalent to **monotonicity of the gradient** $(\nabla f(y) - \nabla f(x))^\top(y - x) \geq 0$. 2oc requires $\nabla^2 f$ to exist at every point and $\mathbf{dom}(f)$ open.

(Convexity preserving operations) $\lambda_i \in \mathbb{R}_+, f_i$ convex, $g : \mathbb{R}^m \rightarrow \mathbb{R}^d$
$f := \max_i f_i \vee f := \sum_i \lambda_i f_i$ convex on $\mathbf{dom}(f) = \cap_i \mathbf{dom}(f_i)$
$g(x) = Ax + b \Rightarrow f(x) = f(g(x))$ convex if f convex on $\{x \in \mathbb{R}^m : g(x) \in \mathbf{dom}(f)\}$

f, g convex $\not\Rightarrow f \circ g$ convex! E.g. $f = -\ln, g = x^2 - 1$, domain will not be convex. f co, g co + non-decreasing $\Rightarrow g(f(x))$ co. f, g co, positive & monotonically incr. $\Rightarrow fg$ co.

(Global minimum) Let f conv, $\mathbf{dom}(f)$ open, $x \in \mathbf{dom}(f)$. Then:

$$x \text{ is global minimum of } f \Leftrightarrow \nabla f(x) = 0$$

$(\Rightarrow)$ doesn't require convexity

If f is **strictly convex**, there is at most one global minimum. $\nabla f(x) > 0\ \forall x \Rightarrow f$ strictly co. $\Leftarrow$: $f(x) = x^4$.

(Constr. opt.) $f : \mathbf{dom}(f) \rightarrow \mathbb{R}$ co+diff. $X \subseteq \mathbf{dom}(f)$ co. $x^* \in X$ is a min $\Leftrightarrow \nabla f(x^*)^\top(x - x^*) \geq 0\ \forall x \in X$.

W'strass: f cont. If sublvl set $f^{\leq\alpha}$ nonempty and bounded, then f has glob min.

Convex programming: $\min f_0(x)$, s.t. $f_i(x) \leq 0, h_j(x) = 0$, $(i = 1..m, j = 1..p)$. Feasible region: $X = \{x \in \mathbb{R}^d : f_i(x) \leq 0, h_j(x) = 0 \forall i, j\}$.

Lagrangian: $L : \mathcal{D} \times R^m \rightarrow \mathbb{R}$, $L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{j=1}^p \nu_j h_j(x)$. λ_i, ν_i are Langrange multipliers.

Dual function: $g : \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R} \cup \{-\infty\}$, $g(\lambda, \nu) = \inf_{x\in D} L(x, \lambda, \nu)$.

Weak duality: If x feasible, then $g(\lambda, \nu) \leq f_0(x)$ for all $\lambda \in \mathbb{R}^m \geq 0, \nu \in \mathbb{R}^p$.

Dual problem: $\max g(\lambda, \nu)$, s.t. $\lambda \geq 0$. Always conv (even if primal isn't).

Slater point: Suppose a conv prog with feasible solution $\tilde{x}$ in addition satisfies $f_i(\tilde{x}) < 0, i = 1..m$ (a Slater point). Then the infimum value of the primal equals the supremum value of the dual. Moreover, if the value is finite, it is attained by a feasible solution of the dual. Note: Strong duality ($\inf f_0(x) = \sup g(\lambda, \nu)$) may also hold when there is no Slater point or even when it's not a conv prog. The stated Slater point condition provides one particular sufficient condition.

KKT conditions: When strong duality holds, KKT provide necessary and –under convexity– sufficient conditions. Let $\tilde{x}, (\tilde{\lambda}, \tilde{\nu})$ be primal and dual optimal solutions with 0 duality gap ($f_0(\tilde{x}) = g(\tilde{\lambda}, \tilde{\nu})$). If all f_i, h_j are differentiable, then (necessary):

$$\tilde{\lambda}_i f_i(\tilde{x}) = 0, \quad i = 1..m$$
$$\nabla f_0(\tilde{x}) + \sum_{i=1}^m \tilde{\lambda}_i \nabla f_i(\tilde{x}) + \sum_{j=1}^p \tilde{\nu}_j \nabla h_j(\tilde{x}) = 0$$

Sufficient: All f_i, h_j diff, all f_i conv, h_j affine and the above equations hold. Then $\tilde{x}, (\tilde{\lambda}, \tilde{\nu})$ have 0 duality gap.

L-smoothness

(L-smoothness) $f : \mathbb{R}^d \rightarrow \mathbb{R}$, conv not req. (!)

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

If f co, the following are equiv.:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$
$$f(y) \geq f(x) + \nabla f(x)^\top(y - x) + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2$$
$$(\nabla f(x) - \nabla f(y))^\top(x - y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2$$
$$(\nabla f(x) - \nabla f(y))^\top(x - y) \leq L\|x - y\|^2$$

Also these: $f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda)f(y) - \frac{\lambda(1-\lambda)L}{2}\|x - y\|^2$ and $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \frac{\lambda(1-\lambda)}{2L}\|\nabla f(x) - \nabla f(y)\|^2$.

For f 2$\times$ diff, also $\nabla^2 f(x) \preceq L\mathbf{I}$ is equiv.

f L-smooth $\Leftrightarrow g(x) := \frac{L}{2}x^\top x - f(x)$ is convex on $\mathbf{dom}(f)$.

All $f(x) = x^\top Qx + b^\top x + c$ are $2\|Q\|$-smooth.

$f = \sum \lambda_i f_i$ is $\sum \lambda_i L_i$-smooth. $f(Ax+b)$ is $L\|A\|^2$-smooth.

μ-strong convexity

(μ-strong convexity) $f : \mathbb{R}^d \rightarrow \mathbb{R}$

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

f μ-sc $\Leftrightarrow g(x) = f(x) - \frac{\mu}{2}x^\top x$ is convex on $\mathbf{dom}(f)$.

f μ-sc $\Leftrightarrow (\nabla f(x) - \nabla f(y))^\top(x - y) \geq \mu\|x - y\|^2$.

f μ-sc $\Leftrightarrow f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \frac{\alpha(1-\alpha)\mu}{2}\|x - y\|^2$.

f μ-sc $\Leftrightarrow \nabla^2 f(x) \succeq \mu\mathbf{I}$.

f μ-sc $\Rightarrow \|\nabla f(x) - \nabla f(y)\| \geq \mu\|x - y\|$.

f μ-sc $\Rightarrow f(y) \leq f(x) + \nabla f(x)^\top(y - x) + \frac{1}{2\mu}\|\nabla f(x) - \nabla f(y)\|^2$.

f μ-sc $\Rightarrow (\nabla f(x) - \nabla f(y))^\top(x - y) \leq \frac{1}{\mu}\|\nabla f(x) - \nabla f(y)\|^2$.

f μ-sc $\Rightarrow f$ strictly convex + has unique global minimum.

f is μ-smooth *and* μ-sc $\Rightarrow f(x) = \frac{\mu}{2}\|x - b\|^2 + c$.

f L-sm and μ-sc $\Rightarrow (\nabla f(x) - \nabla f(y))^\top(x - y) \geq \frac{\mu L}{\mu + L}\|x - y\|^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|^2$.

Convergence

Always w.r.t. $f(x) - f(x^*) < \varepsilon$, as there could be several minima $y^* \neq x^*$. $O(1/\varepsilon)$ better than $O(1/\varepsilon^2)$, but $O(1/T^2)$ better than $O(1/T)$.

Convergence rates (must hold only for sufficiently large t): $\varepsilon_t = f(x_t) - f(x^*)$.

Linear: $\varepsilon_{t+1} \leq c\varepsilon_t, c \in (0,1) \quad \Rightarrow O(\log(1/\varepsilon))$.

Sup.: $\varepsilon_{t+1} \leq c\varepsilon_t^r, c > 0, r > 1; r = 2 \Rightarrow O(\log\log(1/\varepsilon))$.

Sublinear: Anything below linear.

Gradient Descent **(GD)**

$$x_{t+1} = x_t - \gamma\nabla f(x_t)$$

Vanilla analysis: Bound for avg. error since x_T is not necessarily close to best. Result follows from 1oc, UR and cos-thm.

f conv: $\sum_{t=0}^{T-1}\varepsilon_t \leq \frac{\gamma}{2}\sum_{t=0}^{T-1}\|g_t\|^2 + \frac{1}{2\gamma}\|x_0 - x^*\|^2$

f conv, $\|x_0 - x^*\| \leq R, \|\nabla f(x)\| \leq B, \gamma = R/(B\sqrt{T})$: $\frac{1}{T}\sum_{t=0}^{T-1}\varepsilon_t \leq \frac{RB}{\sqrt{T}}$ and $\min_{t=0}^{T-1}\varepsilon_t \leq \varepsilon \Rightarrow T \geq \frac{R^2B^2}{\varepsilon^2}$

(Sufficient decrease) f L-smooth, $\gamma := 1/L$

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}\|\nabla f(x_t)\|^2, \; t \geq 0$$

f conv, L-smooth: $f(x_T) - f(x^*) \leq \frac{L}{2T}\|x_0 - x^*\|^2$ and $T \geq \frac{R^2L}{2\varepsilon}$

f conv, L-sm, μ-sc: vanilla: $\varepsilon_t \leq \frac{1}{2\gamma}(\gamma^2\|\nabla f(x_t)\|^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2) - \frac{\mu}{2}\|x_t - x^*\|^2$. With $\gamma = 1/L$ we get (i) geometrically decr dist to x^* and (ii) exp small abs error after T iter.

$$\|x_{t+1}-x^*\|^2 \le (1-\mu/L)\,\|x_t-x^*\|^2,\ t\ge 0$$

$$f(x_T)-f(x^*) \le \frac{L}{2}(1-\mu/L)^T\,\|x_0-x^*\|^2,\ T>0$$

It follows $T \ge \frac{L}{\mu}\ln\left(\frac{R^2L}{2\varepsilon}\right)$

Projected Gradient Descent (*Proj. GD*)

Choose $x_0\in X$ arb. Proj is well-defined for squared dist, even sc and unique min for closed conv set X.

$$y_{t+1} := x_t-\gamma\nabla f(x_t)$$
$$x_{t+1} := \Pi_X(y_{t+1}) := \arg\min_{x\in X}\|x-y_{t+1}\|^2$$

For $X\subseteq\mathbb{R}^d$ closed and conv, $x\in X, y\in\mathbb{R}^d$, it holds:

- $(x-\Pi_X(y))^\top(y-\Pi_X(y))\le 0$ (angle $\ge 90°$)
- $\|x-\Pi_X(y)\|^2+\|y-\Pi_X(y)\|^2\le\|x-y\|^2$

Proj is **non-expansive**: $\|\Pi_X(x)-\Pi_X(y)\|\le\|x-y\|$.

f co, $X\subseteq\mathbf{dom}(f)$ closed & co, $\|x_0-x^*\|\le R$, $\|\nabla f(x)\|\le B$, $\gamma:=R/(B\sqrt{T})$: $\frac{1}{T}\sum_{t=0}^{T-1}\varepsilon_t\le(RB)/\sqrt{T}$. $\Rightarrow\mathcal{O}(1/\varepsilon^2)$.

f L-sm, $X\subseteq\mathbf{dom}(f)$ closed & co, $\gamma:=1/L$: $f(x_{t+1})\le f(x_t)-\frac{1}{2L}\|\nabla f(x_t)\|^2+\frac{L}{2}\|y_{t+1}-x_{t+1}\|^2$.

f co, L-sm, $X\subseteq\mathbf{dom}(f)$ closed & co, $\gamma:=1/L$: $\varepsilon_t\le\frac{L}{2T}\|x_0-x^*\|^2$.

f co, L-sm, μ-sc, $X\subseteq\mathbf{dom}(f)$ closed & co. With $\gamma:=1/L$ we get (i) geometrically decr dist to x^* and (ii) exp small abs error after T iter. Constrained optimization $\Rightarrow\nabla f(x^*)\neq 0$ possible!

$$\|x_{t+1}-x^*\|^2\le(1-\mu/L)\,\|x_t-x^*\|^2,\ t\ge 0$$

$$\varepsilon_T\le\|\nabla f(x^*)\|\left(1-\frac{\mu}{L}\right)^{T/2}\|x_0-x^*\|+\frac{L}{2}\left(1-\frac{\mu}{L}\right)^T\|x_0-x^*\|^2$$

Coordinate Descent (*CD*)

For GD proved $x_t\to x^*$, here only $f(x_t)\to f(x^*)$.

(PL inequality) f diff w/ glob min x^*. $\exists\mu>0$ s.t.:

$$\frac{1}{2}\|\nabla f(x)\|^2\ge\mu(f(x)-f(x^*)),\quad\forall x\in\mathbb{R}^d$$

μ-sc $\Rightarrow$ PL holds. (PL is a strictly weaker condition, e.g. $f(x_1,x_2)=x_1^2$ satisfies PL but not μ-sc.) Even some non-conv funcs can satisfy PL.

f L-sm, PL holds, $\gamma:=1/L$: $\varepsilon_T\le(1-\mu/L)^T\varepsilon_0$, $T>0$.

(Coord.-wise smooth) f diff, $\mathcal{L}=(L_1,\dots,L_d)\in\mathbb{R}^+_d$. If

$$f(x+\lambda e_i)\le f(x)+\lambda\nabla_i f(x)+\frac{L_i}{2}\lambda^2,\ \forall x\in\mathbb{R}^d,\lambda\in\mathbb{R}$$

holds, cw-sm w/ $\mathcal{L}$. If $L_i=L$, then w/ param L.

Algorithm: Choose $i\in[d]: x_{t+1}:=x_t-\gamma_i\nabla_i f(x_t)e_i$

f $\mathcal{L}$-cw-sm, $\gamma_i=1/L_i$: $f(x_{t+1})\le f(x_t)-\frac{1}{2L_i}|\nabla_i f(x_t)|^2$.

Randomized CD: $i\in[d]$ chosen uniformly at random in step t. f L-sm, PL holds, $\gamma_i=1/L$: $\mathbb{E}[\varepsilon_T]\le(1-\mu/(dL))^T\varepsilon_0, T>0$.

Importance Sampling: choose coordinate actively, sample $i\in[d]$ with prob. $p_i=\frac{L_i}{\sum_{j=1}^d L_j}$. CD-step: $x_{t+1}:=x_t-\frac{1}{L_i}\nabla_i f(x_t)e_i$.

Theorem: f diff with gl. min. x^*. Suppose f cw-sm with param $\mathbb{L}=(L_1,\dots,L_d)$, PL holds with $\mu>0$. Let $\bar{L}=\frac{1}{d}\sum_{i=1}^d L_i$. Then CD with IS and arbitrary x_0 satisfies $\mathbb{E}[f(x_T)-f(x^*)]\le(1-\frac{\mu}{d\bar{L}})^T(f(x_0)-f(x^*)), T>0$.

Steepest CD: $i=\arg\max_i|\nabla_i f(x_t)|$. f L-cw-sm, PL holds, $\gamma_i=1/L$. No $\mathbb{E}$ since alg is deterministic: $\varepsilon_T\le(1-\mu/(dL))^T\varepsilon_0, T>0$. $\Rightarrow$ Difference to GD is that only cw-sm instead of global smoothness is needed. In case f μ-sc wrt ℓ_1-norm (stronger cond.), then d can be dropped in the bound.

Greedy CD: f diff not required. Choose $i\in[d]: x_{t+1}:=\arg\min_{\lambda\in\mathbb{R}}f(x_t-\lambda e_i)$. But now additional 1D opt. problem in each step.

Non-convex functions

f $2\times$ diff, $\|\nabla^2 f(x)\|\le L\,\forall x\in X$. Then f is L-sm.

f L-sm, $\gamma:=1/L$, GD yields: $\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(x_t)\|^2\le\frac{2L}{T}\varepsilon_0$ and $\lim_{t\to\infty}\|\nabla f(x_t)\|^2=0$. Proof using sufficient decr, which doesn't require conv.

Lemma: For f L-sm, GD cannot overshoot a critial point ($\nabla f(x)=0$).

Frank-Wolfe

Constrained opt. $\min_{x\in X}f(x)$

Proj in Proj GD can be expensive even for convex sets.

Linear Min. Oracle: $\mathrm{LMO}_X(g):=\arg\min_{z\in X}g^\top z$.

Algorithm ($\gamma_t\in[0,1]$):

$$s:=\mathrm{LMO}_X(\nabla f(x_t))$$
$$x_{t+1}:=(1-\gamma_t)x_t+\gamma_t s$$

In each step, alg. minimizes the linear approximation over the set X and makes a step in the direction of the minimizer. Iterates are always feasible.

Duality gap / Hearn gap: $g(x):=\nabla f(x)^\top(x-s)$. g can be interpreted as opt gap of the linear subproblem $\nabla f(x)^\top x-\nabla f(x)^\top s$. $g(x)\ge 0$.

Duality gap is an upper bound for the optimality gap: $g(x)\ge f(x)-f(x^*)$. I.e. $g(x_t)$ always gives a guaranteed upper bound on the optimality gap.

f co, L-sm, X closed+bounded, $\mu_t=\gamma_t:=2/(t+2)$, then: $\varepsilon_T\le\frac{2L\mathrm{diam}(X)^2}{T+1}$, $T\ge 1$, $\mathbf{diam}(X):=\max_{x,y\in X}\|x-y\|$.

Descent lemma for $\gamma_t\in[0,1]$: $f(x_{t+1})\le f(x_t)-\gamma_t g(x_t)+\gamma_t^2\frac{L}{2}\|s-x_t\|^2$.

Stepsize variants:

Line search s.t. progress is maximal: $\gamma_t:=\arg\min_{\gamma\in[0,1]}f((1-\gamma)x_t+\gamma s)$. For $h(x)=f(x)-f(x^*)$, we then obtain: $h(x_{t+1})\le h(y_{t+1})\le(1-\mu_t)h(x_t)+\mu_t^2\frac{L}{2}\mathrm{diam}(X)^2$, where y_{t+1} is the iterate obtained using standard stepsize μ_t

Gap-based $\gamma_t:=\min\{1,\frac{g(x_t)}{L\|s-x_t\|^2}\}$ and progress is guaranteed in every iteration: $h(x_{t+1})\le\begin{cases}h(x_t)-(1-\frac{\gamma_t}{2}), & \gamma_t<1\\ h(x_t), & \gamma_t=1\end{cases}$

$(f,X),(f',X')$ **affinely equiv** if $f'(x)=f(Ax+b)$ for A inv. $X'=\{A^{-1}(x-b):x\in X\}$. LMO+FW return same iterates.

Random

Unconstrained optimization:

	Lip+co	L+co	μ+co	L+μ+co
GD	$\mathcal{O}(\varepsilon^{-2})$	$\mathcal{O}(\varepsilon^{-1})$		$\mathcal{O}(\log(\varepsilon^{-1}))$
AGD		$\mathcal{O}(1/\sqrt{\varepsilon})$		
Proj. GD	$\mathcal{O}(\varepsilon^{-2})$	$\mathcal{O}(\varepsilon^{-1})$		$\mathcal{O}(\log(\varepsilon^{-1}))$
Subgr. D	$\mathcal{O}(\varepsilon^{-2})$		$\mathcal{O}(\varepsilon^{-1})$	
SGD	$\mathcal{O}(\varepsilon^{-2})$		$\mathcal{O}(\varepsilon^{-1})$	

LMO: Let $X:=\mathrm{conv}(\mathcal{A})$, then:

Ex.	$\mathcal{A}$	$\lvert\mathcal{A}\rvert$	dim.	$\mathrm{LMO}_X(g)$
L1-ball	$\{\pm e_i\}$	$2d$	d	$\pm e_i, i=\arg\max_i\lvert g_i\rvert$
Simplex	$\{e_i\}$	d	d	$e_i, i=\arg\min_i g_i$
Spectahedron	$\{xx^\top,\lVert x\rVert=1\}$	∞	d^2	$\arg\min_{\lVert x\rVert=1}x^\top Gx$
Norms	$\{x,\lVert x\rVert\le 1\}$	∞	d	$\arg\min_{\lVert s\rVert\le 1}\langle s,g\rangle$
Nuclear norm	$\{Y,\lVert Y\rVert_*\le 1\}$	∞	d^2	..

Performance of AGD vs Subgr. D:

	Convex	Strongly Convex
Subgr. D	$\mathcal{O}\left(\frac{BR}{\sqrt{t}}\right)$	$\mathcal{O}\left(\frac{B^2}{\mu t}\right)$
AGD	$\mathcal{O}\left(\frac{LR^2}{t^2}\right)$	$\mathcal{O}\left(\left(\frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}\right)^{2t}\right)$

$\rightarrow$ Subgr. D is always slower, even in sc case only sublinear cvg.

Complexity for SGD:

	iteration complexity	iteration cost	total
Smooth and strongly convex problems ($\kappa=L/\mu$)			
GD	$\mathcal{O}(\kappa\log(1/\varepsilon))$	$\mathcal{O}(n)$	$\mathcal{O}(n\kappa\log(1/\varepsilon))$
SGD	$\mathcal{O}(1/\varepsilon)$	$\mathcal{O}(1)$	$\mathcal{O}(1/\varepsilon)$
Nonconvex problems			
GD	$\mathcal{O}(1/\varepsilon^2)$	$\mathcal{O}(n)$	$\mathcal{O}(n/\varepsilon^2)$
SGD	$\mathcal{O}(1/\varepsilon^4)$	$\mathcal{O}(1)$	$\mathcal{O}(1/\varepsilon^4)$

Vanilla Analysis (GD & Proj. GD):

1. Use 1oc: $f(y)\ge f(x)+\nabla f(x)^\top(y-x)$
2. Set $y=x^*, x=x_t$: $\varepsilon_t\le\nabla f(x_t)^\top(x_t-x^*)$
3. Use update rule: $x_t-x^*=(z_{t+1}-x^*)+\gamma\nabla f(x_t)$ where $z_{t+1}=x_{t+1}$ for GD, $z_{t+1}=y_{t+1}$ for Proj. GD
4. Apply cosine theorem: $2v^\top w=\|v\|^2+\|w\|^2-\|v-w\|^2$
5. For Proj. GD: Use projection property $\|x_{t+1}-x^*\|^2\le\|y_{t+1}-x^*\|^2$
6. Sum over t, telescope: $\sum_{t=0}^{T-1}\varepsilon_t\le\frac{\gamma}{2}\sum_{t=0}^{T-1}\|\nabla f(x_t)\|^2+\frac{1}{2\gamma}\|x_0-x^*\|^2$

L-smooth:

1. Use smoothness: $f(y)\le f(x)+\nabla f(x)^\top(y-x)+\frac{L}{2}\|y-x\|^2$
2. Set $y=z_{t+1}, x=x_t$, use update rule where $z_{t+1}=x_{t+1}$ for GD, $z_{t+1}=y_{t+1}$ for Proj. GD
3. For Proj. GD: Use projection property $f(x_{t+1})\le f(y_{t+1})$
4. Minimize RHS w.r.t. γ: $\gamma=1/L$
5. Get sufficient decrease:
   GD: $f(x_{t+1})\le f(x_t)-\frac{1}{2L}\|\nabla f(x_t)\|^2$
   Proj. GD: $f(x_{t+1})\le f(x_t)-\frac{1}{2L}\|\nabla f(x_t)\|^2+\frac{L}{2}\|y_{t+1}-x_{t+1}\|^2$

μ-strongly convex:

1. Use strong convexity: $f(y)\ge f(x)+\nabla f(x)^\top(y-x)+\frac{\mu}{2}\|y-x\|^2$
2. Set $y=x^*, x=x_t$, combine with vanilla analysis
3. Use sufficient decrease to bound/eliminate gradient term
4. For Proj. GD: Apply projection property $\|x_{t+1}-x^*\|^2\le\|y_{t+1}-x^*\|^2$
5. Get recursive inequality for $\|x_t-x^*\|^2$

Working with iterate distances: $\|x_{t+1}-x^*\|^2=\|x_t-\gamma\nabla f(x_t)-x^*\|^2=\|x_t-x^*\|^2-2\gamma\nabla f(x_t)^\top(x_t-x^*)+\gamma^2\|\nabla f(x_t)\|^2$ (use update rule and expand norm). Then bound middle term with μ-sc and L-sm or similar properties. For projections in UR: use non-expansive prop.

Telescoping sum: $\sum_{t=0}^{T-1}(f(x_t)-f(x_{t+1}))=f(x_0)-f(x_T)$

Matrix diff example:

$f(x)=\log(a^\top x)\Rightarrow\nabla f(x)=\frac{a}{a^\top x}\Rightarrow\nabla^2 f(x)=-\frac{aa^\top}{(a^\top x)^2}$ ($a_i>0$)

$f(x)=\sum_{i=1}^d\log(x_i)\Rightarrow\nabla f(x)=(\frac{1}{x_1},\dots,\frac{1}{x_d})\Rightarrow\nabla^2 f(x)=-\mathrm{diag}(\frac{1}{x_1^2},\dots,\frac{1}{x_d^2})$

Stochastic: $F(x):=\mathbb{E}_\xi[f_\xi(x)]$. unbiased grad estimator: $\mathbb{E}[\nabla f_\xi(x)]=\nabla F(x)$. Then: $\nabla F(x^*)=\mathbb{E}[\nabla f_\xi(x^*)]=0$. But: $\nabla f_\xi(x^*)\neq 0, \mathbb{E}[\|\nabla f_\xi(x^*)\|^2]\neq 0$. Jensen: $\|\nabla F(x)\|^2=\|\mathbb{E}[\nabla f_\xi(x)]\|^2\le\mathbb{E}[\|\nabla f_\xi(x)\|^2]$

Probability: $\mathbb{E}[X]=\sum x_i p(x_i)$, $\mathrm{Var}[X]=\mathbb{E}[(X-\mathbb{E}[X])^2]=\mathbb{E}[X^2]-\mathbb{E}[X]^2$, $\mathrm{Cov}[X,Y]=\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]=\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]$.

$\mathbb{E}[XY|Z]=\mathbb{E}[X|Z]\mathbb{E}[Y|Z]$ if X,Y indep given Z. $P(B)=\sum P(B|A_i)P(A_i)$, $P(A|B)=\frac{P(B|A)P(A)}{P(B)}$

Newton's method

1D: $x_{t+1} := x_t - f'(x_t)/f''(x_t),\ t \geq 0$.

For optimization apply to f': $x_{t+1} := x_t - f'(x_t)/f''(x_t),\ t \geq 0$, resp. $x_{t+1} := x_t - \nabla^2 f(x_t)^{-1}\nabla f(x_t)$.

f co, 2× diff, $\nabla^2 f(x) \succ 0$ inv, then x_{t+1} from Newton satisfies $x_{t+1} = \arg\min_{x\in\mathbb{R}^d} f(x_t) + \nabla f(x_t)^\top(x - x_t) + \frac{1}{2}(x - x_t)^\top \nabla^2 f(x_t)(x - x_t)$.

Let there be a ball $X \subseteq \mathbf{dom}(f)$ with center x^* such that $\left\|\nabla^2 f(x)^{-1}\right\| \leq 1/\mu$ and $\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\| \leq B\|x - y\|$, then for x_t, x_{t+1} resulting from a Newton step, the following holds: $\|x_{t+1} - x^*\| \leq \frac{2B}{\mu}\|x_t - x^*\|^2$.

f 2× diff, μ-sc over open conv $X \subseteq \mathbf{dom}(f)$. Then $\nabla^2 f(x)$ is inv and $\left\|\nabla^2 f(x)^{-1}\right\| \leq 1/\mu$ for all $x \in X$.

Quasi-Newton methods

Secant method (2nd derivative free!): Replace $f''(x)$ with $\frac{f'(x_t) - f'(x_{t-1})}{x_t - x_{t-1}}$.

Subgradient methods

(Subgradient) $f : \mathbf{dom}(f) \to \mathbb{R}\cup\{+\infty\}$, co. $g \in \mathbb{R}^d$ is a subgradient of f at x if

$$f(y) \geq f(x) + g^\top(y - x),\ \forall y \in \mathbf{dom}(f)$$

Set of all subgradients at x is called subdifferential $\partial f(x)$.

If f co and diff at x, then $\partial f(x) = \{\nabla f(x)\}$.

f co, $\mathbf{dom}(f)$ open, $B \in \mathbb{R}_+$. The following are equiv:

- $\|g\| \leq B,\ \forall x \in \mathbf{dom}(f), \forall g \in \partial f(x)$.
- $|f(x) - f(y)| \leq B\|x - y\|,\ \forall x, y \in \mathbf{dom}(f)$.

If $\mathbf{0} \in \partial f(x), x \in \mathbf{dom}(f)$, then x is a *global* minimum.

f co, $x \in \mathbf{dom}(f)$. Then $\partial f(x)$ is co and closed.

f func where $\mathbf{dom}(f)$ is co and $\partial f(x) \neq \emptyset\ \forall x \in \mathbf{dom}(f)$. Then f is co over $\mathbf{dom}(f)$.

Directional derivatives: $f'(x; d) = \lim_{\delta\to 0^+}\frac{f(x+\delta d) - f(x)}{\delta}$. For f diff $f'(x; d) = \nabla f(x)^\top d$. For subgr: $f'(x; d) = \max_{g\in\partial f(x)} g^\top d$.

Calculating subgradients:

- *Conic combination:* $h(x) = \lambda f(x) + \mu g(x); \lambda, \mu \geq 0; f, g$ co, then $\partial h(x) = \lambda\partial f(x) + \mu\partial g(x)\ \forall x \in \mathbf{int}(\mathbf{dom}(h))$.
- *Affine compos.:* $h(x) = f(Ax + b)$; f co, then $\partial h(x) = A^\top \partial f(Ax + b)$.
- *Supremum:* $h(x) = \sup_{\alpha\in\mathcal{A}} f_\alpha(x)$ and f_α co, then: $\partial h(x) \supseteq \mathbf{conv}\{\partial f_\alpha(x) \mid \alpha \in \alpha(x)\}$ where $\alpha(x) = \{\alpha : h(x) = f_\alpha(x)\}$
- *Superposition:* $h(x) = F(f_1(x), \ldots, f_m(x))$ where $F(y_1, \ldots, y_m)$ is non-decr and co, then $\partial h(x) \supseteq \left\{\sum_{i=1}^m d_i \partial f_i(x) : (d_1, \ldots, d_m) \in \partial F(y_1, \ldots, y_m)\right\}$.

Subgradient method: f co, possibly non-diff. Goal $\min f(x)$ s.t. $x \in X \subseteq \mathbf{dom}(f)$. X closed+co. Let $R^2 = \max_{x,y\in X}\|x - y\|_2^2$, $B = \sup_{x,y\in X}\frac{|f(x) - f(y)|}{\|x - y\|_2}$. Init $x_1 \in X$. For $t = 1, \ldots, T$:

$$x_{t+1} = \Pi_X(x_t - \gamma_t g_t),\ g_t \in \partial f(x_t)$$

For f diff, this reduces to Proj GD. Subgr. Descent is not necessarily a descent method and moving along the negative direction of g_t is not guaranteed to decrease the function value.

Stepsize choices:

- *Constant:* $\gamma_t \equiv \gamma > 0$
- *Scaled:* $\gamma_t = \gamma/\|g_t\|_2$
- *Diminishing, non-summable:* $\sum\gamma_t = \infty, \lim_{t\to\infty}\gamma_t = 0$
- *Sq-summable:* $\sum\gamma_t = \infty, \sum\gamma_t^2 < \infty$ (e.g. $1/t$)
- *Polyak:* Assuming $f(x^*)$ known. $\gamma_t = \varepsilon_t/\|g_t\|_2^2$

f co, then SubgrD satisfies

$$\min \varepsilon_t \leq \left(\sum_{t=1}^T \gamma_t\right)^{-1}\left(\frac{1}{2}\|x_1 - x^*\|_2^2 + \frac{1}{2}\sum_{t=1}^T \gamma_t^2\|g_t\|_2^2\right)$$

$$f(\hat{x}_T) - f(x^*) \leq \left(\sum_{t=1}^T \gamma_t\right)^{-1}\left(\frac{1}{2}\|x_1 - x^*\|_2^2 + \frac{1}{2}\sum_{t=1}^T \gamma_t^2\|g_t\|_2^2\right)$$

where $\hat{x}_T = \left(\sum_{t=1}^T \gamma_t\right)^{-1}\left(\sum_{t=1}^T \gamma_t x_t\right) \in X$.

Using bounds R, B and changing summation to $T_0 \geq 1$:

$$\min_{T_0\leq 1\leq T} f(x_t) - f(x^*) \leq \frac{\frac{R^2}{2} + \frac{1}{2}\sum_{t=T_0}^T \gamma_t^2 B^2}{\sum_{t=T_0}^T \gamma_t}$$

Mirror Descent

Goal: Generalize SubgrD to non-Euclid. distances.

(Bregman divergence) $\omega : X \to \mathbb{R}$ *strictly(!)* conv, continuously diff on closed conv X.

$$V_\omega(x, y) = \omega(x) - \omega(y) - \nabla\omega(y)^\top(x - y)$$

V_ω is not a valid distance: asymmetric and triangle ineq. may not hold–it is called distance-generating function.

If ω σ-sc wrt some norm, then it holds $V_\omega(x, y) \geq \frac{\sigma}{2}\|x - y\|^2$.

For well-defined V_ω, V_ψ and $a, b > 0$ it holds $V_{a\omega+b\psi}(x, y) = aV_\omega(x, y) + bV_\psi(x, y)$.

Generalized Pythagorean: Let x^* be Bregman proj of x_0 onto conv set $C \subset X$, $x^* = \arg\min_{x\in C} V_\omega(x, x_0)$. Then for all $y \in C$: $V_\omega(y, x_0) \geq V_\omega(y, x^*) + V_\omega(x^*, x_0)$.

Prox-mapping: $\mathrm{Prox}_x(\xi) = \arg\min_{u\in X}\{V_\omega(u, x) + \langle\xi, u\rangle\}$, where ω is 1-sc wrt some norm.

Mirror descent:

$$\begin{aligned} x_{t+1} = \mathrm{Prox}_{x_t}(\gamma_t g_t) &= \arg\min_{x\in X}\{V_\omega(x, x_t) + \langle\gamma_t g_t, x\rangle\} \\ &= \arg\min_{x\in X}\{\omega(x) + \langle\gamma_t g_t - \nabla\omega(x_t), x\rangle\} \end{aligned}$$

Example setups

ℓ_2: $X \subseteq R^n, \omega(x) = \frac{1}{2}\|x\|_2^2, \|\cdot\| = \|\cdot\|_2$: $V_\omega(x, y) = \frac{1}{2}\|x - y\|_2^2$; $\mathrm{Prox}_x(\xi) = \Pi_X(x - \xi) \implies$ SubgrD.

ℓ_1: $X = \Delta_n, \omega(x) = \sum_{i=1}^n x_i\ln(x_i), \|\cdot\| = \|\cdot\|_1$: $V_\omega(x, y) = \sum_{i=1}^n x_i\ln(x_i/y_i)$ (Kullback-Leibler); $\mathrm{Prox}_x(\xi) = \left(\sum_{i=1}^n x_i\exp(-\xi_i)\right)^{-1}\left[x_1\exp(-\xi_1), \ldots, x_n\exp(-\xi_n)\right]^\top$ Good for multiplicative updates with normalization.

(Three point iden.) $\forall x, y, z \in \mathbf{dom}(\omega) : V_\omega(x, z) = V_\omega(x, y) + V_\omega(y, z) - \langle\nabla\omega(z) - \nabla\omega(y), x - y\rangle$

Convex conjugate

$f : \mathbf{dom}(f) \to \mathbb{R}$, conv conj: $f^*(y) = \sup_{x\in\mathbf{dom}(f)}\{x^\top y - f(x)\}$. f conv is not necessary!

Fenchel inequality follows from def.: $x^\top y \leq f(x) + f^*(y)$, which is a generalization of Young's ineq $x^\top y \leq \|x\|^2/2 + \|y\|^2/2$.

If f co, lower semi-continuous and proper, then $(f^*)^* = f$. That is $\liminf_{x\to x_0} f(x) \geq f(x_0)$ and $f(x) > -\infty$.

f μ-sc $\implies f^*$ is $1/\mu$-Lipschitz smooth and continuously diff.

For f, g proper, conv, semi-cont:

$$(f + g)^*(x) = \inf_y\{f^*(y) + g^*(x - y)\}$$

$$(\alpha f)^*(x) = \alpha f^*(x/\alpha),\ \alpha > 0$$

Smoothing techniques

Goal: Approximate non-sm/diff f with smooth f_μ s.t. GD and AGD can be applied.

Nesterov's smoothing: $f_\mu(x) = \max_{y\in\mathbf{dom}(f^*)}\{x^\top y - f^*(y) - \mu\cdot d(y)\} = (f + \mu d)^*(x)$, where $d(y)$ is a sc, non-negative proximity function. f_μ is continuously diff and Lipschitz smooth.

Moreau-Yosida: $f_\mu(x) = \min_{y\in\mathbf{dom}(f)}\{f(y) + \frac{1}{2\mu}\|x - y\|_2^2\}$ for $\mu > 0$. It is equiv to Nesterov with $d(y) = \frac{1}{2}\|y\|_2^2$.