Cauchy-Schwarz: $|\mathbf{u}^\top \mathbf{v}| \leq \|\mathbf{u}\|\|\mathbf{v}\|$.

Spectral norm: $\|A\| := \max_{\|\mathbf{v}\|=1} \|A\mathbf{v}\|$.

Mean-value theorem: If $a < b$ and $h : [a, b] \to \mathbb{R}$ continuous and differentiable in (a, b), then there exists $c \in (a, b)$ such that

$$h'(c) = \frac{h(b) - h(a)}{b - a}.$$

Fundamental theorem of calculus: If $a < b$ and h differentiable on an open domain (a, b) and h' continuous on $[a, b]$, then

$$h(b) - h(a) = \int_a^b h'(t)dt.$$

Differentiable: $f : \mathrm{dom}(f) \to \mathbb{R}^m$, where $\mathrm{dom}(f) \subseteq \mathbb{R}^d$ is differentiable at $\mathbf{x}$ if there exists $A \in \mathbb{R}^{m\times d}$ and an error function $r : \mathbb{R}^d \to \mathbb{R}^m$ defined in some neighborhood of $\mathbf{0} \in \mathbb{R}^d$ such that for all $\mathbf{y}$ in the neighborhood of $\mathbf{x}$,

$$f(\mathbf{y}) = f(\mathbf{x}) + A(\mathbf{y} - \mathbf{x}) + r(\mathbf{y} - \mathbf{x}),$$

where

$$\lim_{\mathbf{v}\to\mathbf{0}} \frac{\|r(\mathbf{v})\|}{\|\mathbf{v}\|} = \mathbf{0}.$$

A is then the Jacobian of f at $\mathbf{x}$.

B-Lipschitz: f is B-Lipschitz if

$$\|f(\mathbf{x}) - f(\mathbf{y})\| \leq B\|\mathbf{x} - \mathbf{y}\|,$$

which is equivalent to bounded gradients on open domains (in closed domains, only $\Leftarrow$ holds)

$$\|\nabla f(\mathbf{x})\| \leq B.$$

Cosine theorem: $2\mathbf{v}^\top \mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v} - \mathbf{w}\|^2$.

2 Convexity

Domain must be convex. Strict convexity if inequalities become strict inequalities. Equivalent definitions $\forall \mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$:

- $f(\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda)f(\mathbf{y})$.
- First-order exists: $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x})$.
- First-order exists: $(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))^\top(\mathbf{y} - \mathbf{x}) \geq 0$.
- Second-order exists: $\nabla^2 f(\mathbf{x}) \succeq 0$.

Intuition: f is above its tangential hyperplane at $(\mathbf{x}, f(\mathbf{x}))$.

Jensen's inequality: If f convex, and $\sum_{i=1}^m \lambda_i = 1$, then

$$f\left(\sum_{i=1}^m \lambda_i \mathbf{x}_i\right) \leq \sum_{i=1}^m \lambda_i f(\mathbf{x}_i).$$

The other direction holds for concave functions ($-f$ is convex).

Preserving convexity: Max, sum, and multiplication with positive scalars preserve convexity. $f \circ g$ is convex on $\mathrm{dom}(f \circ g) := \{\mathbf{x} \in \mathbb{R}^m \mid g(\mathbf{x}) \in \mathrm{dom}(f)\}$ if g is affine.

Local minimum: A point $\mathbf{x}$, such that there exists $\epsilon > 0$ with

$$f(\mathbf{x}) \leq f(\mathbf{y}), \quad \forall \mathbf{y} \in \mathrm{dom}(f) \text{ satisfying } \|\mathbf{y} - \mathbf{x}\| < \epsilon.$$

Global minimum: A point $\mathbf{x}$ such that

$$f(\mathbf{x}) \leq f(\mathbf{y}), \quad \forall \mathbf{y} \in \mathrm{dom}(f).$$

If f is convex and differentiable over an open domain, then $\nabla f(\mathbf{x}) = \mathbf{0}$ if and only if $\mathbf{x}$ is a global minimum.

Sublevel set: Let f be continuous (not convex). If there exists a nonempty and bounded sublevel set $f^{\leq\alpha}$, then f has a global minimum.

TODO: Convex programs.

3 Gradient descent

f must be differentiable, then we use the update rule:

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_t \nabla f(\mathbf{x}_t).$$

Vanilla analysis: Assuming only convexity, we get a bound on the summed error

$$\sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f^\star) \leq \frac{\gamma_t}{2}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{2\gamma_t}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

Proof by using first-order convexity on $\mathbf{x}_t$ and $\mathbf{x}^\star$, and rewrite the gradient descent update rule.

Lipschitz functions ($\mathcal{O}(1/\epsilon^2)$): Setting $\gamma := R/B\sqrt{T}$, we get

$$\frac{1}{T}\sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f^\star) \leq \frac{RB}{\sqrt{T}}.$$

Using bound $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$.

3 Smooth functions

L-smooth with equivalent definitions $\forall \mathbf{x}, \mathbf{y} \in X$:

- $f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2$.
- Lemma 3.3: $\frac{L}{2}\mathbf{x}^\top\mathbf{x} - f(\mathbf{x})$ is convex.
- Lemma 3.5: $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$.
- Lemma 6.1: $\|\nabla^2 f(\mathbf{x})\| \leq L$ ($\Leftarrow$ only if X is open).

Intuition: f is below a not-too-steep tangential paraboloid at $(\mathbf{x}, f(\mathbf{x}))$.

Affine functions (Lemma 3.4): $f(\mathbf{x}) = \mathbf{x}^\top Q\mathbf{x} + \mathbf{b}^\top\mathbf{x} + c$ is smooth with parameter $2\|Q\|$ if Q is symmetric.

Sufficient decrease (Lemma 3.7): Choosing $\gamma := 1/L$, gradient descent satisfies

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2, \quad t \geq 0.$$

(Already holds if f is L-smooth over line segment connecting $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$.) Proof by first definition of smoothness, cosine theorem, and gradient descent update rule.

Convergence ($\mathcal{O}(1/\epsilon)$) (Theorem 3.8): Choosing $\gamma := 1/L$, gradient descent yields

$$f(\mathbf{x}_T) - f^\star \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

Proof by starting from vanilla analysis and bounding gradient sum with sufficient decrease.

Accelerated gradient descent achieves $\mathcal{O}(1/\sqrt{\epsilon})$ by using an intermediate variable.

3 Strongly convex functions

μ-strongly convex with equivalent definitions $\forall \mathbf{x}, \mathbf{y} \in X$:

- $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2$.
- Lemma 3.11: $f(\mathbf{x}) - \frac{\mu}{2}\mathbf{x}^\top\mathbf{x}$ is convex.

Intuition: f is above a not-too-flat tangential paraboloid at $(\mathbf{x}, f(\mathbf{x}))$.

Strict convexity (Lemma 3.12): If f is μ-strongly convex, then f is strictly convex.

Geometrically decreasing distances (Theorem 3.14): Choosing $\gamma := 1/L$, gradient descent satisfies

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \left(1 - \frac{\mu}{L}\right)\|\mathbf{x}_t - \mathbf{x}^\star\|^2, \quad t \geq 0.$$

Proof by rewriting vanilla analysis with first definition of strong convexity and sufficient decrease.

Convergence $\mathcal{O}(\log {}^1\!/\epsilon)$ (Theorem 3.14): Choosing $\gamma := {}^1\!/L$, gradient descent yields

$$f(\mathbf{x}_T) - f^\star \leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T \|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

<u>Proof</u> by using geometrically decreasing distances and smoothness with $\nabla f(\mathbf{x}^\star) = \mathbf{0}$.

4 Projected gradient descent

Optimization within closed convex subset $X \subseteq \mathbb{R}^d$.

$$\begin{aligned}\mathbf{y}_{t+1} &:= \mathbf{x}_t - \gamma \nabla f(\mathbf{x}) \\ \mathbf{x}_{t+1} &:= \Pi_X(\mathbf{y}_{t+1}) := \operatorname*{argmin}_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{y}_{y+1}\|^2.\end{aligned}$$

After every step, project back onto X.

Projection properties (Fact 4.1): $\mathbf{x} - \Pi_X(\mathbf{y})$ and $\mathbf{y} - \Pi_X(\mathbf{y})$ form an obtuse angle,

- $(\mathbf{x} - \Pi_X(\mathbf{y}))^\top (\mathbf{y} - \Pi_X(\mathbf{y})) \leq 0.$
- $\|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{y} - \Pi_X(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2.$

Lipschitz functions ($\mathcal{O}({}^1\!/\epsilon^2)$) (Theorem 4.2): Same bound as gradient descent. <u>Proof</u> by replacing $\mathbf{x}_{t+1}$ by $\mathbf{y}_{t+1}$ in the vanilla analysis and using the second projection property with $\mathbf{x} = \mathbf{x}^\star$ and $\mathbf{y} = \mathbf{y}_{t+1}$.

Sufficient decrease (Lemma 4.3): If f is L-smooth, choosing stepsize $\gamma := {}^1\!/L$, we get

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

<u>Proof</u> by the same as gradient descent, but then with projection step.

Smooth functions ($\mathcal{O}({}^1\!/\epsilon)$) (Theorem 4.4): Same result as in gradient descent. <u>Proof</u> by compensating for the extra term in sufficient decrease by the vanilla analysis.

Strongly convex ($\mathcal{O}(\log {}^1\!/\epsilon)$) (Theorem 4.5): Decreasing distances still holds, but extra term in convergence bound when choosing $\gamma := {}^1\!/L$,

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

This is due to the fact that we cannot use $\nabla f(\mathbf{x}^\star) = \mathbf{0}$ in the constrained case.

5 Coordinate descent

Update only one coordinate of $\mathbf{x}_t$ at a time, meaning that we only need to compute the gradient of one coordinate of $\nabla f(\mathbf{x}_t)$.

PL inequality: f has a global minimum $\mathbf{x}^\star$. Definition $\forall \mathbf{x} \in X$:

$$\frac{1}{2}\|\nabla f(\mathbf{x})\|^2 \geq \mu(f(\mathbf{x}) - f(\mathbf{x}^\star)).$$

Strong convexity $\Rightarrow$ PL inequality (Lemma 5.2).

Coordinate-wise smoothness: f is coordinate-wise smooth with $\mathcal{L} = [L_1, \dots, L_d] \in \mathbb{R}^d_+$ if $\forall \mathbf{x}, \mathbf{y} \in X, i \in [d]$:

$$f(\mathbf{x} + \lambda \mathbf{e}_i) \leq f(\mathbf{x}) + \lambda \nabla_i f(\mathbf{x}) + \frac{L_i}{2}\lambda^2.$$

This gives a more fine-grained picture of f than smoothness. It might be the case that all L_i are significantly smaller than the best possible L-smoothness.

Update rule:

$$\begin{aligned}&\text{choose an active coordinate } i \in [d] \\ &\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_i \nabla_i f(\mathbf{x}_t)\mathbf{e}_i.\end{aligned}$$

Coordinate-wise sufficient decrease (Lemma 5.5): With stepsize $\gamma_i = {}^1\!/L_i$, coordinate descent satisfies

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L_i}|\nabla_i f(\mathbf{x}_t)|^2.$$

Randomized coordinate descent convergence (Theorem 5.6): f is coordinate-wise smooth with L and satisfies PL-inequality with μ. Choosing $\gamma_i = \frac{1}{L}$, we get

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

<u>Proof</u> by using coordinate-wise sufficient decrease and taking expectation with respect to i on both sides. Then, expectation over $\mathbf{x}_t$ to remove condition.

Importance sampling convergence (Theorem 5.7): Sample i with probability $L_i/\sum_{j=1}^d L_j$. Let $\bar{L} = {}^1\!/d \sum_{i=1}^d L_i$. Choosing $\gamma_i = {}^1\!/L_i$, we get

$$\mathbb{E}[f(\mathbf{x}_T) - f^\star] \leq \left(1 - \frac{\mu}{d\bar{L}}\right)^T (f(\mathbf{x}_0) - f^\star).$$

<u>Proof</u> by the same method as randomized coordinate descent.

Steepest coordinate descent convergence (Corollary 5.8): Choose index with largest absolute gradient. Same conditions as randomized coordinate descent. Then, we get

$$f(\mathbf{x}_T) - f^\star \leq \left(1 - \frac{\mu}{dL}\right)^T (f(\mathbf{x}_0) - f^\star).$$

TODO: Strong convexity with respect to ℓ_1-norm.

Greedy coordinate descent: Choose the index by one of the above methods, but then perform a line search over that coordinate and minimize by solving a 1-dimensional optimization problem (easy). This does not require f to be differentiable. But, this does not always return the global minimum, since there are functions with points where it can make no progress.

Theorem 5.11: Let f be of the form $f(\mathbf{x}) := g(\mathbf{x}) + h(\mathbf{x})$ with $h(\mathbf{x}) = \sum_i h_i(x_i)$, h_i convex, and g convex and differentiable. If $\mathbf{x}$ is a point that greedy coordinate descent cannot make progress in any coordinate, then $\mathbf{x}$ is a global minimum of f.

6 Nonconvex functions

For nonconvex functions, gradient descent may get stuck in a local minimum, stuck in a saddle point (flat region), or infinitely decrease, but never reach a critical point (e.g. ${}^1\!/e^x$).

Gradient convergence (Theorem 6.2): f is L-smooth. Choosing $\gamma := {}^1\!/L$, then

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{2L}{T}(f(\mathbf{x}_0) - f^\star).$$

In particular, $\|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{2L}{T}(f(\mathbf{x}_0) - f^\star)$ for some $t \in [T-1]$, and $\lim_{t\to\infty}\|\nabla f(\mathbf{x}_t)\|^2 = 0$. This does not mean that it converges to a critical point, since it may never reach a point with 0 gradient, but only move toward it asymptotically. <u>Proof</u> by sufficient decrease, which does not require convexity.

$\gamma := {}^1\!/L$ does not overshoot critical points (Lemma 6.3).

TODO: Trajectory analysis.

7 The Frank-Wolfe algorithm

Constrained optimization algorithm without projection (which can be very complex) by making use of linear minimization oracle:

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

The algorithm is then

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

Reduces non-linear constrained optimization to linear optimization over the same set. Rationale is that the gradient defines the best linear approximation of f at $\mathbf{x}_t$.

Properties: (1) iterates are always feasible, i.e., in X, (2) projection-free, which can be very complex, and (3) iterates have a simple sparse representation, i.e., $\mathbf{x}_t$ is always a convex combination of $\mathbf{x}_0$ and the minimizers $\mathbf{s}_{1:t-1}$.

Let $X = \mathrm{conv}(\mathcal{A})$, then every $\mathbf{s} := \mathrm{LMO}_X(\mathbf{g}) \in \mathrm{conv}(X)$ is a convex combination of atoms, $\mathbf{s} = \sum_{i=1}^n \lambda_i \mathbf{a}_i$ with $\sum_{i=1}^n \lambda_i = 1$. Furthermore, there is always an atom in $\mathcal{A}$ that minimizes the LMO.

ℓ_1**-ball**: The LMO for $X = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_1 \leq 1\}$ is given by

$$\mathrm{LMO}_X(\mathbf{g}) = -\mathrm{sign}(g_i)\mathbf{e}_i \text{ with } i := \operatorname*{argmax}_{i \in [d]} |g_i|.$$

TODO: Spectahedron.

Duality gap (Lemma 7.2): We can easily compute an upper bound of the optimality gap,

$$g(\mathbf{x}) := \nabla f(\mathbf{x})^\top(\mathbf{x} - \mathbf{s}) \geq f(\mathbf{x}) - f^\star,$$

with $\mathbf{s} := \mathrm{LMO}_X(\nabla f(\mathbf{x}))$. At any optimal point $\mathbf{x}^\star$, $g(\mathbf{x}^\star) = 0$. Proof by using $\nabla f(\mathbf{x})^\top \mathbf{s} \leq \nabla f(\mathbf{x})^\top \mathbf{x}^\star$, and the first-order characterization of convexity.

Descent (Lemma 7.4): For a step $\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_t(\mathbf{s} - \mathbf{x}_t)$ with stepsize $\gamma_t \in [0,1]$, it holds that

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \gamma_t f(\mathbf{x}_t) + \gamma_t^2 \frac{L}{2}\|\mathbf{s} - \mathbf{x}_t\|^2,$$

with $\mathbf{s} := \mathrm{LMO}_X(\nabla f(\mathbf{x}))$. Proof by first definition of smoothness and duality gap.

Convergence analysis ($\mathcal{O}(1/\epsilon)$) (Theorem 7.3): f is L-smooth and convex. With $\gamma_t = 2/t+2$, Frank-Wolfe yields

$$f(\mathbf{x}_T) - f^\star \leq \frac{2L\mathrm{diam}(X)^2}{T+1}.$$

Proof by duality gap and descent lemma, and then induction.

Linear search stepsize: Choose $\gamma_t \in [0,1]$ such that the progress is maximized,

$$\gamma_t := \operatorname*{argmin}_{\gamma \in [0,1]} f((1-\gamma)\mathbf{x}_t + \gamma \mathbf{s}).$$

The descent lemma still holds for this stepsize, since this stepsize can only be better than a predetermined stepsize. And, thus the convergence also holds.

TODO: Gap-based stepsize.

TODO: Affine invariance.

TODO: Curvature constant.