10-725 Convex Optimization

Notes by Kadin Zhang, cut from CMU 21-720/721

Contents

1 Convex sets and functions

1.1 Convex sets

Definition (Affine set, hull): A set C is *affine* if for every $x, y \in C$, the line between x and y, $\theta x + (1 - \theta)$, $\theta \in \mathbb{R}$ is in C.

The *affine hull* of a set C is

$$\left\{ \theta_1 x_1 + \cdots + \theta k x_k : \sum_i \theta_i = 1, k = 2, 3, \ldots \right\}.$$

Example: The set of solutions to a system of linear equations $Ax = b$ is affine.

Example (More examples of convex sets):

(a) L_p balls for $p \geq 1$.

(b) Polyhedron: solution set to a finite number of linear inequalities (halfspaces, $a^\top x \leq b$) and equalities (hyperplanes, $a^\top x = b$).

(c) Polytope: bounded polyhedron.

(d) Convex cones: for $x_1, x_2 \in C$, $\theta_1, \theta_2 \geq 0$, $\theta_1 x_1 + \theta_2 x_2 \in C$ (conic combinations are in C).

Definition (Normal cone): Let C be an arbitrary set, $x \in C$. Then

$$N_C(x) = \left\{ g : g^\top (y - x) \leq 0, \forall y \in C \right\}.$$

Definition (Separating hyperplane): A hyperplane $a^\top x = b$ is a *separating hyperplane* for sets C, D if for all $x \in C, y \in D$,

$$a^\top x \leq b, a^\top y \geq b.$$

The separation is *strict* if the inequalities are strict:

$$a^\top x < b, a^\top y > b.$$

The separation is *strong* if there is nonzero margin ε.

Theorem (Separating hyperplane theorems):

(a) If C, D are nonempty disjoint convex sets, then there exists a separating hyperplane between them.

(b) If further the closures do not intersect and one of them is bounded, then there exists a strongly separating hyperplane between them.

Theorem (Supporting hyperplane theorem): Let C convex, and $x_0 \in \partial C$. There exists a hyperplane $a^\top x = b$, $a \neq 0$, such that for all $x \in C$,

$$a^\top x \leq b,$$

and $a^\top x_0 = b$.

Proposition (Convexity preserving set operations):

(a) Affine images and preimages.

(b) Set sum, cross product.

1.2 Convex functions

We will asssume f is defined on the entire space rather than a convex set, since in this case we can define an extended-real-valued $\widetilde{f}$ that maintains convexity and behavior on the original set.

Theorem (Zeroth-order condition / definition): $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if for all $x, y \in \mathbb{R}^n$ and $\theta \in [0, 1]$,

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y).$$

$f : \mathbb{R}^n \to \mathbb{R}$ is convex if for all x in the domain and $v \in \mathbb{R}^n$, $f(x + tv)$ is convex.

Theorem (Epigraph): $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only its epigraph $\{(x,t) : t \geq f(x)\}$ is convex.

Theorem (First order condition): A differentiable $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if for all $x, y \in \mathbb{R}^n$,

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x).$$

Corollary: If f is convex and $\nabla f(x) = 0$, then x is a global minimum.

Definition (Strictly convex): f is *strictly convex* if the inequality in the zeroth order condition is strict for $\theta \in (0, 1)$.

Theorem (Second order condition): A twice differentiable $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if $\nabla^2(x)$ is positive semidefinite for all $x \in \mathbb{R}^n$.

If the hessian matrix is always positive definite, then f is strictly convex. But the reverse is not true, $f(x) = x^4$.

Theorem (Gradients and subgradients monotone): Let $f : \mathbb{R}^n \to \mathbb{R}$ convex. If f is differentiable, then for all $x, y \in \mathbb{R}^n$,

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq 0.$$

This holds for subgradients $g_x \in \partial f(x), g_y \in \partial f(y)$.

Proposition (Convexity preserving operations):

(a) Suppose f_α is convex for $\alpha \in S$. Then $\sup_\alpha f_\alpha(x)$ is convex. Corollary: for arbitrary C, this is convex:

$$f(x) = \max_{y \in C} \|x - y\|$$

(b) If $g(x, y)$ convex and C is convex, then $f(x) = \min_{y \in C} g(x, y)$ is convex. Corollary: for convex C, this is convex:

$$f(x) = \min_{y \in C} \|x - y\|.$$

Proposition (Composition): $h \circ g$ is convex when:

(a) g affine, h convex (or other way around).

(b) g convex, h convex and nondecreasing (easy to see when differentiable).

1.3 Function classes

Assume a domain $\mathbb{Q} \subset \mathbb{R}^d$ for the following function classes.

Definition (Function classes):

(a) $C^k(Q)$ is the set of k-times continuously differentiable functions.

(b) $C^{k,p}_\beta(Q)$ is the set of k-times continuously differentiable functions where the pth derivative is Lipschitz with constant β.

(c) $\mathcal{F}^k(Q)$ is the set of convex k-times continuously differentiable functions.

(d) $\mathcal{F}^{k,p}_\beta(Q)$ is the set of k-times continuously differentiable convex functions where the pth derivative is Lipschitz with constant β.

Definition (β-smooth): The set of β-smooth functions is $C^{1,1}_\beta(Q)$.

Proposition (Smoothness properties):

(a) Let f β-smooth. Then

$$\left| f(y) - f(x) - \nabla f(x)^\top (y - x) \right| \leq \frac{\beta}{2} \|x - y\|^2.$$

(b) Let $f \in C^{2,2}_M$, i.e. $\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\| \leq M\|x - y\|$ and $\|x - y\| = r$. Then,

$$\nabla^2 f(x) - MrI_n \preceq f(y) \preceq \nabla^2 f(x) + MrI_n.$$

(c) If f is β-smooth, then $\frac{\beta}{2}\|x\|^2 - f(x)$ is convex.

Definition (Lipschitz): A function is Lipschitz with constant β if for all x, y,

$$\|f(x) - f(y)\| \le \beta\|x - y\|.$$

Proposition (β-smooth equivalences): The following are equivalent:

(a) ∇f is Lipschitz with constant β;

(b) $(\nabla f(x) - \nabla f(y))^T(x - y) \le \beta\|x - y\|_2^2$ for all x, y;

(c) $\nabla^2 f(x) \preceq \beta I$ for all x;

(d) $f(y) \le f(x) + \nabla f(x)^T(y - x) + \frac{\beta}{2}\|y - x\|_2^2$ for all x, y.

Proof.

- (a) $\implies$ (b). Suppose for all x, y, $\|\nabla f(x) - \nabla f(y)\| \le \beta\|x - y\|$. Then

$$(\nabla f(x) - \nabla f(y))^\top(x - y) \le \|\nabla f(x) - \nabla f(y)\|\|x - y\| \le \beta\|x - y\|^2.$$

- (b) $\implies$ (c). Suppose for all x, y, $(\nabla f(x) - \nabla f(y))^\top(x - y) \le \beta\|x - y\|^2$. Let $\alpha > 0$ be a parameter and consider the hypothesis with points $x, x+\alpha s$, letting $s := y - x$:

$$\begin{aligned} (\nabla f(x + \alpha s) - \nabla f(x))(\alpha s) &\le \beta\|\alpha s\|^2 \\ \implies \frac{1}{\alpha}\left(\int_0^\alpha \nabla^2 f(x + ts)s\,dt\right)s &\le \beta\|s\|^2. \end{aligned}$$

Letting $\alpha \to 0$:

$$s^\top \nabla^2 f(x)s \le \beta\|s\|^2.$$

Since x, $s = y - x$ are arbitrary, this suffices to show that the eigenvalues of the Hessian matrix are at most β, as required.

- (c) $\implies$ (d). Suppose the eigenvalues of $\nabla^2 f$ are at most β. Fix x, y. By Taylor's theorem,

$$f(y) = f(x) + \nabla f(x)^\top(y - x) + \frac{1}{2}(y - x)^\top\nabla^2 f(z)(y - x),$$

for some z between x and y. Using the eigenvalue assumption on the quadratic form yields (d):

$$f(y) \le f(x) + \nabla f(x)^\top(y - x) + \frac{\beta}{2}\|y - x\|^2.$$

- (d) $\implies$ (b). Suppose for all x, y, $f(y) \le f(x)+\nabla f(x)^\top(y-x)+\frac{\beta}{2}\|y - x\|^2$. Using this assumption on x, y, then y, x, adding, and cancelling:

$$0 \le \nabla f(x)^\top(y - x) + \nabla f(y)^\top(x - y) + \beta\|x - y\|^2.$$

Now we can rearrange and use linearity of inner product to get (b).

$$(\nabla f(x) - \nabla f(y))^\top(x - y) \le \beta\|x - y\|^2.$$

- (c) $\implies$ (a). Suppose the eigenvalues of $\nabla^2 f$ are bounded by β. By mean value theorem we have for any x, y

  $$\|\nabla f(x) - \nabla f(y)\| = \left\|\nabla^2 f(z)(x-y)\right\|,$$

  for some z between x and y. Applying our eigenvalue assumption on $\nabla^2 f$ yields (a):

  $$\|\nabla f(x) - \nabla f(y)\| \leq \beta\|x-y\|.$$

□

Proposition (Strong convexity):

(a) A function f is α-strongly convex if $g(x) = f(x) - \frac{\alpha}{2}\|x\|_2^2$ is convex (definition).

(b) f is α-strongly convex if and only if for all $\theta \in [0,1]$,

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

Proposition (Strong convexity equivalences): The following are equivalent:

(i) f is strongly convex with constant α;

(ii) $(\nabla f(x) - \nabla f(y))^T(x-y) \geq \alpha\|x-y\|_2^2$ for all x, y;

(iii) $\nabla^2 f(x) \succeq \alpha I$ for all x;

(iv) $f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\alpha}{2}\|y-x\|_2^2$ for all x, y.

Proof. Similar to Lipschitz equivalences. □

Definition (Strictly convex): A function f is *strictly convex* if for $x \neq y$,

$$f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta)f(y).$$

2 Matrices

2.1 SVD

Definition (Singular value decomposition): Let X be $n \times m$. We can decompose

$$X = U\Sigma V^\top,$$

where U contains the left singular vectors, V contains right singular vectors, and Σ is the diagonal matrix with singular values.

X is a mapping $\mathbb{R}^m \to \mathbb{R}^n$. In analog to $Ax = \lambda x$, for SVD we have

$$Xv_i = \sigma_i u_i.$$

In direct analog with eigenvalues, $\sigma_1(A) = \max_{x \in S^{n-1}} \|Ax\|$.

2.2 Norms

Note that norms are convex by triangle inequality and homogeneity:

$$\|\theta x + (1-\theta)y\| \leq \theta\|x\| + (1-\theta)\|y\|.$$

Definition (Operator norm): Let A be a $m \times n$ matrix. We define the operator norm as the largest factor by which the linear operator A can stretch a vector:

$$\|A\| := \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{x \in S^{n-1}} \|Ax\|_2.$$

Note that this is equal to the largest singular value of A.

Definition (Induced norm): The operator norm is a special case of an induced p-norm, $1 \leq p \leq \infty$,

$$\|A\|_p := \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}.$$

Example:

(a) $\|A\|_1$ is the maximum absolute column sum of A,

$$\|A\|_1 = \max_j \sum_{i=1}^m \|A_{ij}\|.$$

(b) $\|A\|_\infty$ is the maximum absolute row sum of A,

$$\|A\|_1 = \max_i \sum_{j=1}^n \|A_{ij}\|.$$

Definition (Nuclear/Trace norm):

$$\|A\|_{Tr} := \operatorname{tr}\left(\sqrt{A^\top A}\right) = \sum \sigma_i(A).$$

Definition (Frobenius norm):

$$\|A\|_F := \sqrt{\sum_i \sum_j |A_{ij}|^2} = \sqrt{\mathrm{tr}(A^\top A)} = \sqrt{\sum \sigma_i(A)^2}.$$

Proposition (Spectral and Frobenius norms are monotone): The spectral and Frobenius norms are monotone: if $A \preceq B$, then $\|A\| \leq \|B\|$.

Definition (Schatten p-norms):

$$\|A\|_{Schp} := \left(\sum_i \sigma_i(A)^p \right)^{1/p}.$$

Note that $p = \infty$ yields the spectral norm, $p = 2$ yields the Frobenius norm, and $p = 1$ yields the nuclear norm.

3 Gradient descent

3.1 Gradient descent

Proposition (Gradient descent is regularized linear approximation): Suppose we perform gradient descent update with update parameter η, i.e.

$$x^{t+1} = x^t - \eta \nabla f(x^t).$$

This is equivalent to

$$x^{t+1} = \arg\min_y f(x^t) + \nabla f(x^t)^\top (y - x^t) + \frac{1}{2\eta} \left\| y - x^t \right\|^2.$$

Definition (Convergence rates): Let s_i be a sequence with limit s and

$$0 \le \delta = \frac{|s_{i+1} - s|}{|s_i - s|} \le 1.$$

(a) s_i has *linear convergence* if $\delta \in (0, 1)$.

(b) s_i has *superlinear convergence* if $\delta = 0$.

(c) s_i has *sublinear convergence* if $\delta = 1$.

If further

$$\lim_{n \to \infty} \frac{|s_{n+1} - s|}{|s_n - s|^2} < \infty,$$

then s_n has *quadratic convergence*.

Theorem (Least squares gradient descent convergence rate): Suppose we are solving a least squares problem

$$\hat{x} = \arg\min_x \frac{1}{2} \|Ax - b\|^2.$$

Suppose $S := A^\top A$ has finite condition number

$$\kappa(S) = \frac{\lambda_{\max}(S)}{\lambda_{\min}(S)}.$$

If we fix $\eta = \frac{2}{\lambda_{\max}(S) + \lambda_{\min}(S)}$, then we have linear convergence:

$$\left\| x^k - \hat{x} \right\| \le \left(\frac{\kappa(S) - 1}{\kappa(S) + 1} \right)^k \left\| x^0 - \hat{x} \right\|.$$

Theorem (GD on β-smooth functions): Let f be β-smooth. Then for fixed step size $\eta \leq 2/\beta$, GD is a descent algorithm. For $\eta \leq 1/\beta$,

$$f(x^{t+1}) \leq f(x^t) - \frac{\eta}{2}\left\|f(x^t)\right\|^2.$$

If f is further twice differentiable, and x^* is minimizer of f, then with $\eta = \frac{1}{\beta}$, the gradient will exhibit sublinear convergence:

$$\min_{i \leq k}\left\|\nabla f(x^k)\right\| \leq O\left(\frac{\sqrt{\beta}}{\sqrt{k}}\right).$$

If f is further convex, then

$$\min_{i \leq k} f(x^k) - f(x^*) \leq O\left(\frac{\beta}{k}\right).$$

Theorem (GD on β-smooth and α-strongly convex functions): Let f β-smooth and α-strongly convex. Let x^* be a minimizer of f. Define condition number $\kappa = \beta/\alpha > 1$. Then,

$$\min_{i \leq k} f(x^k) - f(x^*) \leq O\big(c^k\big),$$

where $c = 1 - \frac{1}{\kappa}$.

3.2 Subgradient method

Definition (Subgradient, subdifferential): Let $f : \mathbb{R}^n \to \mathbb{R}$ convex. A *subgradient* of f at x is g_x such that for all y,

$$f(y) \geq f(x) + g_x^\top (y - x).$$

The set of all subgradients at x is called the *subdifferential*

$$\partial f(x) = \{g \in \mathbb{R}^n : g \text{ is subgradient of } f \text{ at } x\}.$$

Definition (Normal cone): Let C be a set, $x \in C$. The normal cone of C at x is

$$N_C(x) \coloneqq \big\{g : g^\top (y - x) \leq 0, \forall y \in C\big\}.$$

Definition (Subgradient method): Repeat

$$x^{k+1} = x^k - \eta_k g^k,$$

where g^k is any subgradient of f at x^k.

Example (Subgradient method isn't descent): Let $f(x_1, x_2) = |x_1| + 10|x_2|$. At $(1, 0)$, $(1, 10)$ is a valid subgradient, so the next point could be $(1 - \eta, -10\eta)$ which yields strictly larger f.

Proposition: If we choose subgradient with minimum norm, this is a descent direction.

Proposition: If f is G-Lipschitz continuous, $\|g\| \le G$ for all subgradients g of f.

Lemma: Let f be G-Lipschitz and convex. Let $\eta_0, \ldots, \eta_k$ be a sequence of step sizes. Let $\hat{x}$ be the best iterate among $x^0, \ldots, x^k$. The subgradient method satisfies

$$f(\hat{x}) - f(x^*) \le \frac{\|x^* - x_0\|^2 + G^2 \sum_{i=0}^{k-1} \eta_i^2}{2 \sum_{i=0}^{k-1} \eta_i}$$

Theorem (Convergence rate with constant step size): Let f be G-Lipschitz and convex. Let $\hat{x}$ be the best iterate among $x^0, \ldots, x^k$. Let $\|x^0 - x^*\| \le R$. Choose constant step size $\eta = \frac{R}{G\sqrt{k}}$. Then, subgradient method satisfies

$$f(\hat{x}) \le f(x^*) + \frac{GR}{\sqrt{k}}.$$

Theorem (Convergence with diminishing step size): Let f be G-Lipschitz and convex. Let $\hat{x}$ be the best iterate among $x^0, \ldots, x^k$. Let $\|x^0 - x^*\| \le R$. Choose step size

$$\sum_i \eta_i^2 < \infty, \sum_i \eta_i = \infty.$$

Then $\lim_{k\to\infty} f(\hat{x}) = f(x^*)$.

3.3 Projected subgradient method

Definition (Projected subgradient method): Let C be a convex set and f be a convex function. Suppose we want to solve

$$\min_{x \in C} f(x).$$

Projected subgradient method iterates

$$\begin{aligned} y^{t+1} &= x^t - \eta_t g_{x^t} \\ x^{t+1} &= P_C(y^{t+1}), \end{aligned}$$

where $g_{x^t} \in \partial f(x^t)$ and $P_C(y^{t+1}) = \arg\min_{x \in C} \|x - y^{t+1}\|$.

Proposition:

Theorem (Convergence for projected subgradient method): Let $f : \mathbb{R}^d \to \mathbb{R}$ convex and G-Lipschitz, $\eta_0, \ldots, \eta_k$ step sizes. Let $\hat{x}$ be the closest iterate among $x^0, \ldots, x^k$. Then,

$$f(\hat{x}) - f(x^*) \leq \frac{\|x^* - x_0\|^2 + G^2 \sum_{i=0}^{k-1} \eta_i^2}{2 \sum_{i=0}^{k-1} \eta_i}$$

Proof. Essentially same as non-projected case but using the fact that projection onto C is a contraction, i.e. for all x, y,

$$\|P_C(x) - P_C(y)\| \leq \|x - y\|.$$

□

3.4 Optimality conditions

In the unconstrained case,

Theorem: x^* is optimal if and only if $0 \in \partial f(x^*)$.

Proof. $0 \in \partial f(x^*) \iff$ for all y,

$$f(y) \geq f(x^*) + 0^\top (y - x^*) = f(x).$$

□

Theorem (Constraint optimality condition, differentiable case): Let $f : \mathbb{R}^n \to \mathbb{R}$ convex and C convex. A point x is optimal in $\min_{x \in C} f(x)$ is optimal if and only if the negative gradient at x belongs to $N_C(x)$, i.e. there is no decrease direction within the set.

Proof. Let $-\nabla f(x^*) \in N_C(x^*)$. Then $-\nabla f(x^*)^\top (y - x^*) \leq 0$ for all $y \in C$, thus

$$\nabla f(x^*)^\top (y - x^*) \geq 0,$$

and convexity gives us $f(y) \geq f(x^*) + \nabla f(x^*)(y - x^*) \geq f(x^*)$.

On the other hand, assume x^* is optimal but there is y such that $\nabla f(x^*)(y - x^*) < 0$. Then f is locally decreasing on the segment from x^* to y (can formalize this by defining a function $h(t) = f(x^* + t(y - x^*))$ and showing it has negative derivative), a contradiction. □

Theorem (Non-differentiable case): x^* is optimal if and onlly if $0 \in \partial f(x^*) + N_C(x^*)$.

Example (Optimality condition for projection): Let $x^* = P_C(y)$ be a solution to $\min_{x \in C} \frac{1}{2}\|y - x\|^2$. Then for all $c \in C$,

$$(y - x^*)^\top (c - x^*) \leq 0.$$

Proof. By optimality condition, $-\nabla f(x^*) = y - x^* \in N_C(x^*)$. So for all $c \in C$, by normal cone,

$$(y - x^*)^\top (c - x^*) \leq 0.$$

□

Corollary (Projection is a contraction): Let C convex. Then,

$$\|P_C(x) - P_C(y)\| \leq \|x - y\|.$$

3.5 Stochastic gradient descent

Definition (Stochastic gradient): Let $g(x; \xi)$ depend on some randomness ξ such that

$$\mathbf{E} g(x; \xi) = \nabla f(x).$$

Then g is the *stochastic gradient*, and *stochastic gradient descent* follows update rules

$$x^{t+1} = x^t - \eta_t g(x^t; \xi^t).$$

Definition (Empirical risk minimization): Say we're trying to calculate

$$f^* = \arg\min_f \mathbf{E} \ell(f(X), Y),$$

where expectation is taken with respect to X and Y. Then

Example:

(a) Mini-batch SGD: we can take

Theorem (SGD convergence rate, fixed step size): With fixed step size $\eta < 1/\alpha$, we have

$$O(c^k + C)$$

convergence where $c < 1$, $C > 0$.

Theorem (SGD convergence rate, decaying step size): Suppose f is convex, $\mathbf{E} g(x; \xi) \in \partial f(x)$. Let $\left\|x^0 - x^*\right\|^2 \leq R$. Suppose $\mathbf{E}\|g(x; \xi)\|^2 \leq G^2$ (i.e. f is G-Lipschitz?). Then for $\eta = \frac{\sqrt{R}}{\sqrt{G^2 k}}$,

$$\mathbf{E} f\left(\frac{1}{k} \sum_{t=1}^{k} x^t\right) - f(x^*) \leq \frac{G\sqrt{R}}{\sqrt{k}}.$$

Note that this attains the same convergence rate as the subgradient method, but in expectation. We can prove high probability analogues with more difficulty. However, SGD will converge much faster as the iterations can be less expensive.

Theorem (SGD convergence rate, strong convexity): Suppose f is α-strongly convex and that $\mathbf{E}\|g(x;\xi)\|^2 \le G^2$ and $\mathbf{E}g(x;\xi) = \nabla f(x)$ (we can prove similar result for subgradient). For fixed step size $\eta < 1/\alpha$,

$$\mathbf{E}\left\|x^k - x^*\right\|^2 \le (1-\alpha\eta)^k \left\|x^0 - x^*\right\|^2 + \frac{\eta G}{\alpha}.$$

For $\eta_t = \frac{1}{\alpha(t+1)}$,

$$\mathbf{E}f\left(\frac{1}{k}\sum_{t=1}^{k} x^t\right) - f(x^*) \le \frac{G^2(1+\log k)}{2\alpha k}.$$

3.6 Proximal gradient descent

Suppose we have a composite function of the form

$$f = g + h,$$

where g is convex and differentiable, and h is convex but non-smooth.

Definition (Proximal operator):

$$\operatorname{prox}_{\eta,h}(\tilde{x}) = \arg\min_z \frac{1}{2\eta}\|z - \tilde{x}\|^2 + h(z).$$

Definition (Proximal gradient descent):

$$x^{t+1} = \operatorname{prox}_{\eta_t,h}(x^t - \eta_t \nabla g(x^t)).$$

We can also define a *generalized gradient*

$$G_\eta(x) = \frac{x - \operatorname{prox}_\eta(x - \nabla g(x))}{\eta},$$

so that the updates are

$$x^{t+1} = x^t - \eta_t G_{\eta_t}(x^t).$$

Theorem (Proximal operator for Lasso): Let

$$S_\lambda(x) = \arg\min_z \frac{1}{2}\|x - z\|^2 + \lambda\|z\|_1.$$

Then

$$[S_\lambda(x)]_i = \begin{cases} x_i - \lambda, & \text{if } x_i > \lambda; \\ 0, & \text{if } |x_i| \le \lambda; \\ x_i + \lambda, & \text{otherwise.} \end{cases}$$

Proof. We solve this problem for each entry:

$$\min_{z_i} \frac{1}{2}(x_i - z_i)^2 + \lambda |z_i|.$$

The objective is not smooth, so we need

$$0 = z_i - x_i + \lambda u_i,$$

where $u_i \in \partial |z_i|$.

If $z_i > 0$, then $u_i = 1$, and $z_i = x_i - \lambda$. If $z_i < 0$, then $z_i = x_i + \lambda$. If $z_i = 0$, then $x_i = \lambda u_i$, where $u_i \in [-1, 1]$, so

$$x_i \in [-\lambda, \lambda].$$

□

Theorem (Proximal GD descent lemma): Let f β-smooth and $\eta \leq 1/\beta$. For any z,

$$f(x - \eta G_\eta(x)) \leq f(z) + G_\eta(x)^\top (x - z) - \frac{\eta}{2} \|G_\eta(x)\|^2.$$

If g is further α-strongly convex, then

$$f(x - \eta G_\eta(x)) \leq f(z) + G_\eta(x)^\top (x - z) - \frac{\eta}{2} \|G_\eta(x)\|^2 - \frac{\alpha}{2} \|x - z\|^2.$$

Proof, α-strongly convex case. First we bound $g(x - \eta G_\eta(x))$ and $f(x - \eta G_\eta(x))$. By smoothness of g,

$$g(x - \eta G_\eta(x)) \leq g(x) + \nabla g(x)^\top (-\eta G_\eta(x)) + \frac{\beta}{2} \|\eta G_\eta(x)\|^2.$$

By strong convexity g,

$$g(z) \geq g(x) + \nabla g(x)^\top (z - x) + \frac{\alpha}{2} \|z - x\|^2.$$

Combining,

$$g(x - \eta G_\eta(x)) \leq g(z) + \nabla g(x)^\top (x - z) + \nabla g(x)^\top (-\eta G_\eta(x)) + \frac{\beta}{2} \|\eta G_\eta(x)\|^2 - \frac{\alpha}{2} \|x - z\|^2$$

On the other hand, by convexity h,

$$h(x - \eta G_\eta(x)) \leq h(z) - \partial h(x - \eta G_\eta(x))^\top (z - (x - \eta G_\eta(x))).$$

By proposition from lecture,

$$G_\eta(x) - \nabla g(x) \in \partial h(x - \eta G_\eta(x)).$$

Plugging this in as a subgradient,

$$h(x - \eta G_\eta(x)) \leq h(z) - (G_\eta(x) - \nabla g(x))^\top (z - x) - \eta \|G_\eta(x)\|^2 + \eta \nabla g(x)^\top G_\eta(x).$$

Finally we add both inequalities, using $\eta\beta \leq 1$:

$$\begin{aligned}
&f(x - \eta G_\eta(x)) \\
&= g(x - \eta G_\eta(x)) + h(x - \eta G_\eta(x)) \\
&\leq g(z) + \nabla g(x)^\top (x - z) + \nabla g(x)^\top (-\eta G_\eta(x)) + \frac{\beta}{2}\|\eta G_\eta(x)\|^2 - \frac{\alpha}{2}\|x - z\|^2 \\
&+ h(z) - (G_\eta(x) - \nabla g(x))^\top (z - x) - \eta\|G_\eta(x)\|^2 + \eta\nabla g(x)^\top G_\eta(x) \\
&\leq f(z) + \left(\frac{\beta\eta^2}{2} - \eta\right)\|G_\eta(x)\|^2 + G_\eta(x)(x - z) - \frac{\alpha}{2}\|x - z\|^2 \\
&= f(z) + G_\eta(x)(x - z) - \frac{\eta}{2}\|G_\eta(x)\|^2 - \frac{\alpha}{2}\|x - z\|^2.
\end{aligned}$$

□

Theorem (Proximal GD convergence rate): Let f β-smooth and $\eta = \frac{1}{\beta}$.

$$\left\|f(x^k) - f(x^*)\right\| \leq \frac{\beta\left\|x^0 - x^*\right\|^2}{2k}.$$

Note that this is a $1/k$ rate of convergence for a class of non smooth functions, compared to $1/\sqrt{k}$ rate for subgradient method on arbitrary non smooth functions.

Proof, α-strongly convex case. First we plug $z = x^*$, $x = x^t$ into the descent lemma from (1):

$$f(x^{t+1}) \leq f(x^*) + G_\eta(x^t)^\top (x^t - x^*) - \frac{\eta}{2}\left\|G_\eta(x^t)\right\|^2 - \frac{\alpha}{2}\left\|x^t - x^*\right\|^2.$$

Rearranging:

$$\frac{\eta}{2}\left\|G_\eta(x^t)\right\|^2 - G_\eta(x^t)^\top (x^t - x^*) \leq f(x^*) - f(x^{t+1}) - \frac{\alpha}{2}\left\|x^t - x^*\right\|^2. \quad (*)$$

Now,

$$\begin{aligned}
\left\|x^{t+1} - x^*\right\|^2 &= \left\|x^t - \eta G_\eta(x^t) - x^*\right\|^2 \\
&= \left\|x^t - x^*\right\|^2 - 2\eta(x^t - x^*)^\top G_\eta(x^t) + \eta^2\left\|G_\eta(x^t)\right\|^2 \\
&\leq \left\|x^t - x^*\right\|^2 + 2\eta(f(x^*) - f(x^{t+1}) - \frac{\alpha}{2}\left\|x^t - x^*\right\|^2) && (*) \\
&\leq \left\|x^t - x^*\right\|^2 - 2\eta\frac{\alpha}{2}\left\|x^t - x^*\right\|^2 && (f(x^*) - f(x^{t+1}) \leq 0) \\
&= (1 - 1/\kappa)\left\|x^t - x^*\right\|^2. && (\eta = \tfrac{1}{\beta})
\end{aligned}$$

It follows that for any $k \geq 0$,

$$\|x^k - x^*\|_2^2 \leq (1 - 1/\kappa)^k \|x^0 - x^*\|_2^2,$$

as required. □

Example (Matrix completion (Soft impute)): Suppose we have a set of indices I of a matrix for which we know Y_{ij} for $(i,j) \in I$. We wish to find a matrix M that "fills in" the other values while staying as low rank as possible:

$$M^* = \arg\min_M \sum_{(i,j)\in I} (Y_{ij} - M_{ij})^2 + \lambda \operatorname{rank}(M).$$

rank is not convex, so we take its convex envelope (largest convex function lower bounding) over the set of $m \times n$ matrices with norm at most 1, the Nuclear norm $\|X\|_* = \sum_i \sigma_i(X)$. Note that this indeed lower bounds the rank, since the rank gives us the number of nonzero singular values, each of which is bounded by 1. So we solve

$$M^* = \arg\min_M \sum_{(i,j)\in I} (Y_{ij} - M_{ij})^2 + \lambda\|M\|_*.$$

Similar to in l1 norm, we can show that

$$\operatorname{prox}_{\lambda,h}(M) = U\Sigma_\lambda V^\top,$$

where $U\Sigma V^\top = M$ and $\Sigma_\lambda = \max(0, \Sigma - \lambda)$.

4 Duality

4.1 Linear programs

Definition (Linear program):

$$\begin{aligned} \min \ & c^\top x \\ \text{s.t. } & Ax = b \\ & Gx \le h \end{aligned}$$

where $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, $G \in \mathbb{R}^{r\times n}$.

Geometrically, we have m equality constraints and r inequality constraints on x, so the feasible set is a polyhedron in $\mathbb{R}^n$. We slide the normal hyperplane to c from the "negative direction", and look for the first point this hyperplane intersects with the feasible polyhedron.

Note that we can equivalently present this as

$$\begin{aligned} \min_x \ & c^\top x \\ \text{s.t. } & a_i^\top x = b_i, \quad i = 1, \dots, m \\ & x \ge 0 \end{aligned}$$

since each inequality constraint $g_i^\top x \le h_i$ is equivalent to $g_i^\top x + y_i = h_i$ with $y_i \ge 0$.

Definition (Dual linear program): The goal is to lower bound $c^\top x$ for feasible x. Let $u \in \mathbb{R}^m, v \in \mathbb{R}^r$ with $v \ge 0$. Then,

$$u^\top (Ax - b) + v^\top (Gx - h) \le 0.$$

Combining x terms:

$$(-u^\top A - v^\top G)x \ge -u^\top b - v^\top h.$$

From this we have the dual formulation:

$$\begin{aligned} \max_{u,v} \ & -b^\top u - h^\top v \\ \text{s.t. } & -A^\top u - G^\top v = c \\ & v \ge 0 \end{aligned}$$

4.2 Lagrangian

Definition (General form for minimization problems and Lagrangian):

$$\begin{aligned} \min_x \ & f(x) \\ \text{s.t. } & h_i(x) \le 0 \quad i = 1, \dots, m \\ & \ell_j(x) = 0 \quad j = 1, \dots, r \end{aligned}$$

Then, the *Lagrangian function* is

$$L(x, u, v) = f(x) + \sum_j u_j \ell(x) + \sum_i v_i h_i(x).$$

Definition (Lagrange dual function): Let C be the feasible primal region. Then,

$$\min_{x \in C} f(x) \geq \min_{x \in C} L(x, u, v) \geq \min_{x \in \mathbb{R}^n} L(x, u, v) := g(u, v).$$

The dual problem is

$$\max_{u,v} \; g(u, v)$$
$$\text{s.t. } v \geq 0$$

Proposition (Dual function is concave): $g(u, v)$ is the pointwise minimum of a set of affine functions in u, v, which is concave.

4.3 Strong duality

Theorem (Slater's theorem): Suppose there is a feasible point x that is strictly feasible for all non-affine inequality constraints. Then $p^* = d^*$.

Proposition: For arbitrary real valued f,

$$\sup_y \inf_x f(x, y) \leq \inf_x \sup_y f(x, y).$$

As a corollary, we have weak duality:

$$p^* = \inf_{x \in C} f(x) = \inf_{x \in C} \sup_{u,v} L(x, u, v) \geq \sup_{u,v} \inf_{x \in \mathbb{R}^n} L(x, u, v) = d^*.$$

Note that we define the Lagrangian to be ∞ for $x \notin C$.

Definition (Saddle point): A *saddle point* of the Lagrangian $L(x, u, v)$ is point such that for all x, u, v,

$$L(x^*, u, v) \leq L(x^*, u^*, v^*) \leq L(x, u^*, v^*).$$

Theorem: A point is a *saddle point* if and only if strong duality holds.

Theorem (KKT point): Let f, h convex and ℓ affine. Assume all are differentiable. Then $(\hat{x}, \hat{u}, \hat{v})$ is a KKT-point if $\hat{x}$ is primal feasible, $\hat{v} \geq 0$ (dual feasible), $\hat{v}_i h_i(\hat{x}) = 0$ for all $i = 1, \ldots, m$, and $\frac{\partial L}{\partial x} = 0$:

$$\nabla f(\hat{x}) + \sum_i \hat{v}_i \nabla h_i(\hat{x}) + \sum_j \hat{u}_j \nabla \ell_j(\hat{x}) = 0.$$

Theorem (KKT sufficiency for convex problems): Let f, h convex, ℓ affine. Suppose $(\hat{x}, \hat{u}, \hat{v})$ satisfy KKT conditions. Then $\hat{x}$ is primal optimal, $(\hat{u}, \hat{v})$ is dual optimal, and strong duality holds.

Proof. From the stationary condition, we have $\nabla_x L(\hat{x}, \hat{u}, \hat{v}) = 0$. Since L is convex in x, it follows that $\hat{x}$ minimizes $L(x, \hat{u}, \hat{v})$. So

$$g(u, v) = \min_x L(x, u, v) \implies g(\hat{u}, \hat{v}) = L(\hat{x}, \hat{u}, \hat{v}).$$

By $\ell_i = 0$ (primal feasibility) and $v_i h_i = 0$ (complementary slackness), we also have $L(\hat{x}, \hat{u}, \hat{v}) = f(\hat{x})$. So

$$g(\hat{u}, \hat{v}) = L(\hat{x}, \hat{u}, \hat{v}) = f(\hat{x}).$$

Now, from weak duality, optimality and thus strong duality follow:

$$g(\hat{u}, \hat{v}) = f(\hat{x}) \geq g(u, v), \;\; f(\hat{x}) = g(\hat{u}, \hat{v}) \leq f(x).$$

□

Theorem (KKT necessity with strong duality): Suppose strong duality holds, x^* is primal optimal, (u^*, v^*) is dual optimal. Then (x^*, u^*, v^*) is a KKT point.

Proof. Primal and dual feasibility come from the assumptions. For complementary slackness,

$$\begin{aligned} f(x^*) &= g(u^*, v^*) \\ &= \inf_x L(x, u^*, v^*) \\ &\leq f(x^*) + \sum_i v_i^* h_i(x^*) + \sum_j u_j^* \ell_j(x^*) \\ &= f(x^*) + \sum_i v_i^* h_i(x^*) && \text{(primal feasibility)} \\ &\leq f(x^*). && \text{(primal, dual feasibility)} \end{aligned}$$

So the inequalities are equalities, and thus $v_i^* h_i(x^*) = 0$. For the stationary condition, since $f(x^*) = \inf_x L(x, u^*, v^*)$, and L is convex in f, we must have

$$\nabla_x L(x^*, u^*, v^*) = 0.$$

□

4.4 Semidefinite programming

Recall the alternative linear program formulation

$$\begin{aligned} \min_x \;\; & c^\top x \\ \text{s.t.} \;\; & a_i^\top x = b_i, \; i = 1, \ldots, m \\ & x \geq 0 \end{aligned}$$

and dual

$$\begin{aligned} \max_{y,s} \;\; & y^\top b \\ \text{s.t.} \;\; & y^\top A + s^\top = c^\top \\ & s \geq 0 \end{aligned}$$

Definition (SDP):

$$\begin{aligned} \min_{X} \ & \langle C, X\rangle \\ \text{s.t. } & \langle A_i, X\rangle = b_i, \ i = 1, \ldots, m \\ & X \succeq 0 \end{aligned}$$

where $C, X, A_i \in S^n$ and $b_i \in \mathbb{R}$. The dual is

$$\begin{aligned} \max_{y,S} \ & y^\top b \\ \text{s.t. } & S = C - \sum_{i=1}^m y_i A_i \\ & S \succeq 0 \end{aligned}$$

where $y, b \in \mathbb{R}^m$ and $S \in S^n$.

Alternatively we can define the primal as

$$\begin{aligned} \min_{x} \ & c^\top x \\ \text{s.t. } & F_0 + x_1 F_1 + \cdots + x_m F_m \succeq 0 \end{aligned}$$

where $x, c \in \mathbb{R}^m$ and $F_i \in S^n$. From this formulation it is easy to see that the feasible region is convex.

Proposition (Duality gap):

$$\begin{aligned} \langle C, X\rangle - y^\top b &= \langle C, X\rangle - \sum_{i=1}^m y_i b_i \\ &= \langle C, X\rangle - \sum_{i=1}^m y_i \langle A_i, X\rangle \\ &= \left\langle C - \sum_{i=1}^m y_i A_i, X \right\rangle. \end{aligned}$$

Since the inner product of two PSD matrices is nonnegative, we have weak duality. We can show that if this gap is 0, then X is primal optimal and y, S are dual optimal.

Example: We may construct SDP with:

(a) $\inf_x c^\top x$ finite, but no feasible x attains it (in the linear case it would be attainable).

(b) Positive duality gap.

Theorem (Strong duality for SDP): Suppose there is strictly primal feasible $X \succ 0$ and strictly dual feasible $S \succ 0$. Then strong duality holds and both optimum can be achieved.

5 Newton's method

Definition (Newton's method): Consider the first order Taylor approximation of the gradient:

$$\nabla f(x+d) \approx \nabla f(x) + \nabla^2 f(x) d.$$

Newton's method updates to the root of this approximation at each step:

$$0 = \nabla f(x) + \nabla^2 f(x) d \implies d = -(\nabla^2 f(x))^{-1} \nabla f(x).$$

Remark: Alternatively, consider the second order Taylor approximation at x_k:

$$\begin{aligned} &f(x) \approx f(x_k) + \nabla f(x_k)^\top (x - x_k) + \frac{1}{2}(x - x_k) \nabla^2 f(x_k)(x - x_k) \\ \implies &\nabla f(x) = \nabla f(x_k) + \nabla^2 f(x_k)(x - x_k). \end{aligned}$$

This also recovers the same step size as we set the gradient to 0. Note that intuitively, Newton's method

Remark: Note that the eigenvalues and eigenvectors of the Hessian give us the curvature in different directions. So the update rule

$$x_{t+1} = x_t - \eta (\nabla^2 f(x_t))^{-1} \nabla f(x_t)$$

says that we should decrease the update in the directions with greatest curvature (to avoid overshooting) and increase the update in directions with least curvature (to speed up in flat regions).

Theorem (Convergence rate): Suppose f has continuous L-Lipschitz second derivative, there is local minimum x^* of f with $f''(x^*) \succeq \ell I$ for $\ell > 0$, and x_0 is close enough to x^* (?).

Then with Newton's method, $|x_n - x^*|$ exhibits quadratic convergence.

6 Matrix factorization

6.1 ICA

Definition (PCA): Suppose we have a $n \times m$ data matrix X with mean zero rows. Consider projecting entries onto a unit vector v in the feature space $\mathbb{R}^m$. The length of projections are Xv, and the sample variance of lengths is $\frac{1}{n}\|Xv\|$. Maximizing over v yields variance $\sigma_1(X)$ with the corresponding right singular vector.

Definition (ICA): Suppose we have $x = Ay$, where y represents independent signals and A is a mixture. We want to recover y. ICA does this by

(a) Remove mean, $\mathbf{E}x = 0$.

(b) Whitening, $\mathbf{E}xx^\top = I$.

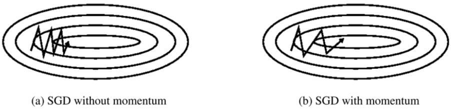
(c) Optimize over orthogonal $W = A^{-1}$ using Jacobi rotation matrices:

$$W = \arg\min_{\tilde{W} \in \mathcal{W}} \mathcal{J}(\tilde{W}x),$$

where $\mathcal{J}$ is a measure of dependence (Shannon mutual information, Kurtosis).

7 Momentum based optimization

Definition (Polyak's momentum method): We add a constant multiple of the previous step direction onto the gradient at the current step: $\theta_{t+1} = \theta_t - v_t$, where $v_t = \gamma v_{t-1} + \eta \nabla F(\theta_t)$.


(a) SGD without momentum (b) SGD with momentum

This especially outperforms standard gradient descent when the gradient is much steeper in one direction than another (poor conditioning).

Definition (Nesterov's accelerated gradient): Polyak's momentum method, but we take the gradient at the location with the previous step: $v_t = \gamma v_{t-1} + \eta \nabla F(\theta_t - \gamma v_{t-1})$.

Definition (AdaGrad): We approximate second order information in the Newton's update

$$\theta_{t+1} = \theta_t - \alpha(\nabla^2 F(\theta_t))^{-1} \nabla F(\theta_t),$$

with a diagonal matrix H_t whose ith entry is

$$\sqrt{\nabla F(\theta_1)^2 + \cdots + \nabla F(\theta_t)^2}.$$

This is an approximation of curvature: dimensions where the gradient has been small in past timesteps should have low curvature, where dimensions where the gradient has been large should have high curvature.

Definition (RMSprop): Let $g_{t,i}$ be the ith coordinate of the gradient, and let $h_{t,i}$ be a exponentially decaying moving average of the squared gradients, $h_{t,i} = \gamma h_{t-1,i} + (1-\gamma) g_{t,i}^2$. Then, define $RMS[g]_{t,i} = \sqrt{h_{t,i} + \varepsilon}$. RMSprop performs updates

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{RMS[g]_{t,i}} g_{t,i}.$$

Remark: RMSprop is unitless. AdaDelta addresses this by setting

$$\theta_{t+1,i} = \theta_{t,i} - \frac{RMS[\Delta\theta]_{t,i}}{RMS[g]_{t,i}} g_{t,i}.$$

Definition (Adam): Combine RMSprop and momentum. Define

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1-\beta_1) g_t \\ v_t &= \beta_2 v_{t-1} + (1-\beta_2) g_t^2. \end{aligned}$$

Note that these are biased: $\mathbf{E} m_t = (1-\beta_1^t)\mathbf{E} g_t$ (proof omitted), so we correct with $\hat{m}_t, \hat{v}_t$. Define update

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t + \varepsilon}} \hat{m}_t.$$