LPs
Standard Form:
$\min c^T x$ s.t. $Ax = b, x \geq 0, b \geq 0$.
Getting it to standard form:
Getting rid of $\geq, \leq$:
$x_1 \leq 4 \to x_1 + x_2 = 4, x_2 \geq 0$
Getting rid of $-$ vars:
$x \in \mathbb{R} \to x = u - v, u, v \in \mathbb{R}^+$
Bounded vars:
$x \in [2,5] \to 2 \leq x, x \leq 5$.

Simplex algorithm:
(1) Take cost function, turn into $\min z$ s.t. $c^T x = z$, remainder in standard LP form.
(2) Pivoting: do Gaussian Elimination to get rid of as many variables as possible, without distributing the z around.
(3) Variables that have been eliminated except in one equation are dependent/basic; others independent/non-basic. Can always get a feasible point by setting non-basic variables to zero, and reading out basic variables.

$$\begin{bmatrix} 1 & 0 & C \\ 0 & I_m & A \end{bmatrix} [-z, x_B, x_N]^T = [-z_0, b]^T$$

(4) Improve solutions: find smallest reduced cost C_j. If $C_J \geq 0$, optimality reached, quit. Else, J is incoming.
(5) Find as far as we can go by picking outgoing variable:
$r = \operatorname{argmin}_{i|A_{i,j}>0} b_i/A_{i,j}$
(6) Perform elimination to get rid of J, using equation that makes the outgoing variable a basic one. That is, take the only equation in which the outgoing variable is non-zero, and eliminate the incoming variable with it.
(7) Repeat from 4 until optimality reached.

Convex sets,fcns:
Defns:
A set is is X if for any weighted sum of data points satisfying Y, the weighted sum is in the set.
Convex: $\sum_i \theta_i = 1$, $\theta_i \geq 0$
Affine: $\sum_i \theta_i = 1$.
Conic: $\theta_i \geq 0$.
Examples:
Lines, line segments, hyperplanes, halfspaces, L_p balls for $p \geq 1$, polyhedrons, polytopes.
Preserving operations:
Translation, scaling, intersection, Affine functions (e.g., projection, coordinate dropping), set sum $\{c_1 + c_2 | c_1 \in C_1, c_2 \in C_2\}$, direct sum $\{(c_1, c_2) | c_1 \in C_1, c_2 \in C_2\}$, perspective projection.
Conv. Fcn. Defn:
$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$
$f(y) \geq f(x) + \nabla f(x)^T (y - x)$
Preserving operations, functions:
Non-negative weighted sum, pointwise-max, affine map $f(Ax+b)$, composition, perspective map.

Strict, Strong Convexity
Defns:
Strict convexity:
$f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta) f(y)$ (basically, not linear).
m-Strong convexity:
$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y) - \frac{1}{2} m\theta(1-\theta)||x-y||_2^2$$
Better strong convexity defns:
$(\nabla f(x) - \nabla f(y))^T (x - y) \geq m||x-y||_2^2$
$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{m}{2}||y-x||_2^2$
$\nabla^2 f(x) \geq mI$.

Gradient Descent
Given x^0, repeat $x^k = x^{k-1} - t_k \nabla f(x^{k-1})$.
Picking t: can diverge if t too big, too slow if t too small.
Backtracing line search: start with $t = 1$, while $f(x - t\nabla f(x)) > f(x) - \alpha t||\nabla f(x)||_2^2$, update $t = \beta t$ with $0 < \alpha < 1/2$, $0 < \beta < 1$.

Subgradients
Defn.:
Subgradient of convex f is g s.t.
$f(y) \geq f(x) + g^T(y - x)$
Subdifferential $\partial f(X)$: set of all g.
SG calculus:
$\partial(af) = a\partial f$; $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$; $\partial f(Ax+b) = A^T \partial f(Ax+b)$.
Finite-pointwise max: $\partial \max_{f \in F} f(x)$ is the convex hull of the active (achieving max functions at x).
Norms: if $f(x) = ||x||_p$ and $1/p + 1/q = 1$, then $||x||_p = \max_{||z||_q \leq 1} z^T x$; thus $\partial ||x||_p = \{y : ||y||_q \leq 1, y^T x = \max_{||z||_q \leq 1} z^T x\}$.
Optimality: $f(x^*) = \min f(x) \leftrightarrow 0 \in \partial f(x^*)$
Remember that sgs may not exist for non-convex functions!

Subgradient Method
Given x^0, repeat $x^k = x^{k-1} - t_k g^{k-1}$
SG method not descent method; keep track of best so far.
Picking t: square summable but not summable (e.g., $1/t$). Polyak steps: $(f(x^{k-1}) - f(x^*))/||g^{k-1}||_2^2$.
Projected sg method: Project after taking a step.

Generalized GD
Suppose $f(x) = g(x) + h(x)$ with g convex, diff, h convex, not necessarily diff.
Define $\operatorname{prox}_t(x) = \operatorname{argmin}_z \frac{1}{2t}||x - z||_2^2 + h(z)$; GGD is:
$x^k = \operatorname{prox}_t(x^{k-1} - t_k \nabla g(x^{k-1}))$
Generalized gradient since if
$G_t(x) = (1/t)(x - \operatorname{prox}_t(x - t\nabla g(x)))$
then update is
$x^k = x^{k-1} - t_k G_t(x^{k-1})$
With backtracking: While $g(x - tG_t(x)) > g(x) - t\nabla g(x)^T G_t(x) + \frac{t}{2}||G_t(x)||_2^2$ (maybe with α in last term?) update $t = \beta t$.

Example (Lasso): Prox is $\operatorname{argmin}_z \frac{1}{2t}||\beta - z||_2^2 + \lambda||z||_1 = S_{\lambda t}(\beta)$. $S_\lambda(\beta)$ is the soft-threshold operator,
$$[S_\lambda(\beta)]_i = \begin{cases} \beta_i - \lambda & : \beta_i > \lambda \\ 0 & : -\lambda \leq \beta_i \leq \lambda \\ \beta_i + \lambda & : \beta_i < -\lambda \end{cases}$$

Example (Matrix Completion): Objective: $\frac{1}{2}\sum_{(i,j) \text{ observ}} (Y_{i,j} - B_{i,j})^2 + \lambda||B||_*$ with $||B||_* = \sum_{i=1}^r \sigma_i(B)$.
Prox function: $\operatorname{argmin}_Z \frac{1}{2t}||B - Z||_F^2 + \lambda||Z|_*$.
Solution: matrix soft-thresholding; $U\Sigma_\lambda V^T$ where $B = U\Sigma V^T$ and $(\Sigma_\lambda)_{ii} = \max\{\Sigma_{ii} - \lambda, 0\}$.

Newton's Method: Originally developed for finding roots; use it to find roots of gradient. Want $\nabla f(x) + \nabla^2 f(x)\Delta_x = 0$; solution is $\Delta_x = -[\nabla^2 f(x)]^{-1} \nabla f(x)$.
Damped Newton method:
$x^{k+1} = x^k - h_k[\nabla^2 f(x)]^{-1} \nabla f(x)$.

Conjugate Direction methods: Want to solve $\min \frac{1}{2} x^T Q x - b^T x$ with $Q > 0$. Define Q-orthogonality as $d_i^T Q d_j = 0$.
Exp. subspace thm.:
Let $\{d_i\}_{i=0}^{n-1}$ be Q-conjugate.
(for method) $g_k = Qx_k - b$
$x_{k+1} = x_k + \alpha d_k$
$\alpha_k = -g_k^T d_k / (d_k^T Q d_k)$
Proof sketch ($g_k \perp B_k$) by ind.:
$g_{k+1} = Qx_{k+1} - b = Q(x_k + \alpha_k d_k) - b$
$(Qx_k - b) + \alpha Q d_k = g_k + \alpha Q d_k$
From here, by defn of α, $d_k^T g_{k+1} = d_k^T(g_k + \alpha Q d_k) = d_k^T g_k - \alpha d_k^T Q d_k = 0$
Algorithm:
Arbitrary x_0, repeat $d_0 = -g_0 = b - Qx_0$
$\alpha_k = -g_k^T d_k / d_k^T Q d_k$; $x_{k+1} = x_k + \alpha_k d_k$
$g_k = Qx_k - b$; $d_{k+1} = -g_{k+1} + \beta_k d_k$
$\beta_k = g_{k+1}^T Q d_k / (d_k Q d_k)$

Quasi-Newton Methods:
Gist: approximate Hessian/inverse Hessian.
Symmetric rank-one correction:
Update: $x_{k+1} = x_k - \alpha H_k g_k$
$\alpha_k = \operatorname{argmin}_\alpha f(x_k - \alpha H_k g_k)$ (LS)
$g_k = \nabla f_k$
$H_{k+1} = H_k + \frac{(p_k - H_k q_k)(p_k - H_k q_k)^T}{q_k^T (p_k - H_k q_k)}$
$p_k = x_{k+1} - x_k$; $q_k = g_{k+1} - g_k$
Might not be PSD!
DFP (Rank 2)
$$H_{k+1} = H_k + \frac{p_k p_k^T}{p_k^T q_k} - \frac{H_k q_k q_k^T H_k}{q_k^T H_k q_k}$$
BFGS
Update inverse of Hessian via Sherman-Morrison).
Let $q_k = g_{k+1} - g_k$
$$H_{k+1} = H_k + (1 + \frac{q_k^T H_k q_k}{p_k^T q_k}) \frac{p_k p_k^T}{p_k^T q_k} - \frac{p_k q_k^T H_k + H_k q_k p_k^T}{q_k p_k}$$

LP Duality
Let c_n, $A_{m \times n}$, b_m, $G_{r \times n}$, h_r.
(P) $\min c^T x$ s.t.
$Ax = b$, $Gx \leq h$
(D) $\max -b^T u - h^T v$ s.t.
$-A^T u - G^T v = c$, $v \geq 0$.

Duality:
Consider $\min f(x)$ s.t.
$h_i(x) \leq 0$, $i = 1, \ldots, m$
$l_j(x) = 0$ $j = 1, \ldots, r$
Lagrangian:
$L(x, u, v) = f(x) + \sum_{i=1}^m u_i h_i(x) + \sum_{j=1}^r v_j l_j(x)$ with $u \in \mathbb{R}^m$, $v \in \mathbb{R}^r$ and $u \geq 0$.
Note: $f(x) \geq L(x, u, v)$ at feasible x.
Dual problem:
Let $g(u, v) = \min_x L(x, u, v)$. Lagrange dual function is g. Dual problem $\max_{u \geq 0, v} g(u, v)$.
Note: dual problem always concave.
Strong duality:
Always have $f^* \geq g*$ where $f*$, $g*$ primal and dual objectives. When $f^* = g^*$, have strong duality. If primal is a convex problem (f, h_i convex, l_j affine) and exists a strictly feasible x, then strong duality.

Dual example (lasso):
Have primal:

$\min_\beta \frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1$;
Introduce dummy z and solve:
$\min_{\beta,z} \frac{1}{2}||y - z||_2^2 + \lambda||\beta||_1$ s.t. $z = X\beta$.
Dual is then:
$\min_{\beta,z} \frac{1}{2}||y - z||_2^2 + \lambda||\beta||_1 + u^T(z - X\beta)$
$\frac{1}{2}||y||_2^2 - \frac{1}{2}||y - u||_2^2 - I_{v:||v||_\infty \leq 1}(X^T u/\lambda)$
Or $\min_u \frac{1}{2}\left(||y||_2^2 - ||y - u||_2^2\right)$ s.t. $||X^T u||_\infty \leq \lambda$.

KKT Conditions:
Stationarity:
$0 \in \partial f(x) + \sum_{i=1}^m u_i \partial h_i(x) + \sum_{j=1}^r \partial l_j(x)$
Complementary slackness:
$u_i \cdot h_i(x) = 0$ for all i
P feas.: $h_i(x) \leq 0$, $l_j(x) = 0$ for all i, j
D feas.: $u_i \geq 0$ for all i *Necessary:* if strong duality, then if x^*, u^*, v^* solutions, then they satisfy KKT conditions.
Sufficient: always, if x^*, u^*, v^* satisfy KKT, then primal dual solutions.
Correspondence Under strong duality, x^* achieves the minimum in $L(x, u^*, v^*)$; if $L(x, u^*, v^*)$ has a unique minimum, then the corresponding point is the primal solution.

Correspondence, Conjugates:
Defn. convex conjugate: Given f, $f^*(y) = \max_x y^T x - f(x)$.
Implies $f(x) + f^*(y) \geq x^T y$. If f closed and convex, $^{**} = f$.

Example, norm:
If $f(x) = ||x||$, $f^*(y) = I_{z:||z||_* \leq 1}(y)$

Ellipsoid method for LP: Solves feasibility problems, but any LP can be turned into a feasibility problem. *Setup:* Let Ω be the set satisfying the constraints. Assume $\Omega \subseteq R$-radius ball centered at y_0, and there is a ball with radius r centered at y^* inside Ω. We know R, r, y_0, but not y^*. *Iterations:* Can check if center of ellipsoid ϵ_k is in Ω; if so, done. Else: find a constraint that is violated, find side that is not violated, fit ellipsoid to that half.
Convergence:

$$\frac{\mathrm{Vol}(\epsilon_k)}{\mathrm{Vol}(\epsilon_0)} \leq \left(\frac{\tau}{R}\right)^m \leq \left(\frac{1}{2}\right)^{k/m}$$

which implies $k \leq O(m^2 \log R/\tau)$ where $\tau = 1/(m+1)$.

Penalty Methods:
Original constrained problem (P), $\min_{x \in S} f(x)$, replace with unconstrained problem $\min f(x) + cp(x)$. p satisfies: p continuous, $p(x) \geq 0$, $p(x) = 0$ iff $x \in S$.
Idea: find some solution, increasingly penalize outside S by increasing $c \to \infty$:
Penalty functions:
$p(x) = \frac{1}{2}\sum_{i=1}^p \max([0, g_i(x)])^2$

Barrier Methods:
Replace original problem with $\min_x f(x) + \frac{1}{c}B(x)$ where B is continuous; $B(x) \geq 0$ for all $x \in \mathrm{int}(S)$; $B(x) \to \infty$ as $x \to \partial S$.
Idea: start out in interior, don't let the algorithm leave S. Increase $c \to \infty$. *Barrier functions:*
Suppose $g_i(x) \leq 0$:
$B(x) = -\sum_{i=1}^m \frac{1}{g_i(x)}$
$B(x) = -\sum_{i=1}^m \log(-g_i(x))$

SDP: Inner product: $\mathrm{tr}(A \cdot B) = \sum\sum A_{i,j}B_{i,j}$

ICA: Step 1: whiten. Step 2: want to minimize gaussian-likeness. But non-convex and lots of local minima. Assume additive linear model.
Whitening: $\Sigma = cov(X) = UDU^T$, $A^* = D^{-1/2}U^T A$.
Coordinate descent: Do argmin on each dimension, updating one-by-one. When does coordinate descent work? $g(x) + \sum_i h_i(x_i)$
Non-convex problems: Specialized approach for each.

Convex Conjugates:

$$f^*(y) = \max_x x^T x^* - f(x)$$
$$- \min_f(x) - x^T x^*$$

$f(ax)$	$f^*(x^*/a)$
$f(x+b)$	$f^*(x^*) - b^T x^*$
$af(x)$	$af^*(x^*/a)$
e^x	$x^*\log(x^*) - x*$
$\|\|x\|\|$	$I_{\|\|z\|\|_* \leq 1}(x^*)$

Matrix derivatives:

$$\begin{aligned} \partial A &= 0 \\ \partial(aX) &= a\partial X \\ \partial(\mathrm{tr}(X)) &= \mathrm{tr}(\partial X) \\ \partial(XY) &= (\partial X)Y + X(\partial Y) \\ \partial x^T a/\partial x &= a \\ \partial x^T X b/\partial X &= ab^T \end{aligned}$$

Suppose s,r are functions of x and A is constant,

$$\frac{\partial s^T A r}{\partial x} = \frac{\partial s}{\partial x}^T A r + \frac{\partial r}{\partial x}^T A^T s$$

Matrix properties:
SVD: $A = U\Sigma V^T$ where:
U are the eigenvectors of AA^T
$D = \sqrt{\mathrm{diag}(\mathrm{eig}(AA^T))}$
V are the eigenvectors of $A^T A$.
Can also write A as the weighted sum of r rank-1 matrices. The rank-1 matrices are $\Sigma_{ii}U_iV_i^T$ for $1 \leq i \leq r$.

EVD: $X = VDV^{-1}$ with D diagonal. If X is symmetric, $VV^T = I$.

Traces: Linear.
$\mathrm{tr}(A) = \mathrm{tr}(A^T)$
$\mathrm{tr}(X^T Y) = \mathrm{tr}(XY^T)$
$\mathrm{tr}(X^T Y) = \mathrm{vec}(X)^T \mathrm{vec}(Y)$
$\mathrm{tr}(ABC) = \mathrm{tr}(BCA) = \mathrm{tr}(CAB)$
P^{-1} exists, $\mathrm{tr}(A) = \mathrm{tr}(P^{-1}AP)$.
$\mathrm{tr}(A) = \sum_i \lambda_i$
Sherman-Morrison Mat. Inv.: Suppose A^{-1} exists, $1 + v^T A^{-1} u \neq 0$.
$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^T A^{-1}}{1 + v^T A^{-1} u}$
Matrix norms:
Trace/Nuclear norm:
$||A||_* = \sum_{i=1}^r \sigma_i(a)$
Spectral/Operator norm:
$||A||_{op} = \sigma_1(A)$
Frobenius norm:
$||A||_F = \mathrm{tr}(A^T A)$.

Derivatives:

$f(x)g(x)$	$f'(x)g(x) + f(x)g'(x)$
$f(g(x))$	$f'(g(x))g'(x)$
x^n	nx^{n-1}
$1/f(x)$	$-f^{-2}f'(x)$
$f(x)/g(x)$	$(f'(x)g(x) - g'(x)f(x))/(g(x)^2)$
e^x	e^x
$\ln(x)$	$1/x$
$\log_c(x)$	$1/(x\ln(c))$

Miscellaneous math:
Lipschitz: A function f is Lipschitz continuous if $|f(x_1) - f(x_2)| \leq L|x_1 - x_2|$; controls how quickly the function changes.
Gradient Lipschtiz:
A differentiable function f has Lipschitz continuous gradient $||\nabla f(y) - \nabla f(x)|| \leq L||y - x||$; if it is twice-differentiable, $LI \geq \nabla^2 f(x)$.
Useful inequalities:
Cauchy-Schwarz: $|x^T y| \leq ||x|| \cdot ||y||$.
Hölder: $||fg||_1 \leq ||f||_p||g||_q$ for $1/p + 1/q = 1$.

	Gr.	SG.	Prox.	New.	Conj.	QN	Bar.	P/D IPM
Crit	f sm	any	sm g + simple h	$2\times$ sm	$2\times$	$2\times$	$2\times$	$2\times$
Const.	Proj.	Proj.	Const. Prox	Equality	None	None	$2\times$ sm. ineq.	$2\times$ sm. ineq.
Param.	fix t/LS	$t \to 0$	fix t/LS	fix $t = 1$/LS	fix/LS	LS	in: fixed/LS; out.: bar. $\to \infty$	in:LS out.: bar. $\to \infty$
Cost/It.	chp	chp	? prox	Exp. (∇^2)	$\approx$ chp	$\approx$ chp +Storage	V.Exp	$\approx$ Exp
Rate	$O(1/\epsilon)$	$O(1/\epsilon^2)$	$O(1/\epsilon)$	$O(\log(\log(1/\epsilon)))$	super-lin.	superlin.	$O(\log(1/\epsilon))$	$O(\log(1/\epsilon))$

Gr. and Prox. Gr. are $O(1/\sqrt{\epsilon})$ w/ accel., $O(\log(1/\epsilon))$ w/strong convexity.