Definition 41 [Subgradient]: *Let $f : \mathcal{D} \rightarrow \Re$ be a convex function defined on a convex set $\mathcal{D}$. A vector $\mathbf{h} \in \Re^n$ is said to be a subgradient of f at the point $\mathbf{x} \in \mathcal{D}$ if*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{h}^T(\mathbf{y} - \mathbf{x})$$

for all $\mathbf{y} \in \mathcal{D}$. The set of all such vectors is called the subdifferential of f at $\mathbf{x}$.

Theorem 76 *Let $f : \mathcal{D} \rightarrow \Re$ be a convex function defined on a convex set $\mathcal{D}$. A point $\mathbf{x} \in \mathcal{D}$ corresponds to a minimum if and only if*

$$\nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) \geq 0$$

for all $\mathbf{y} \in \mathcal{D}$.

If $\nabla f(\mathbf{x})$ is nonzero, it defines a supporting hyperplane to $\mathcal{D}$ at the point $\mathbf{x}$. Theorem 77 implies that for a differentiable convex function defined on an open set, every critical point must be a point of (global) minimum.

Theorem 77 *Let $f : \mathcal{D} \rightarrow \Re$ be differentiable and convex on an open convex domain $\mathcal{D} \subseteq \Re^n$. Then $\mathbf{x}$ is a critical point of f if and only if it is a (global) minimum.*

Theorem 78 *Let $f : \mathcal{D} \rightarrow \Re$ with $\mathcal{D} \subseteq \Re^n$ be differentiable on the convex set $\mathcal{D}$. Then,*

1. *f is convex on $\mathcal{D}$ if and only if is its gradient ∇f is monotone. That is, for all $\mathbf{x}, \mathbf{y} \in \Re$*

$$\left(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right)^T (\mathbf{x} - \mathbf{y}) \geq 0 \tag{4.53}$$

2. *f is strictly convex on $\mathcal{D}$ if and only if is its gradient ∇f is strictly monotone. That is, for all $\mathbf{x}, \mathbf{y} \in \Re$ with $\mathbf{x} \neq \mathbf{y}$,*

$$\left(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right)^T (\mathbf{x} - \mathbf{y}) > 0 \tag{4.54}$$

3. *f is uniformly or strongly convex on $\mathcal{D}$ if and only if is its gradient ∇f is uniformly monotone. That is, for all $\mathbf{x}, \mathbf{y} \in \Re$,*

$$\left(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right)^T (\mathbf{x} - \mathbf{y}) \geq c||\mathbf{x} - \mathbf{y}||^2 \tag{4.55}$$

for some constant $c > 0$.

Necessity: Suppose f is uniformly convex on $\mathcal{D}$. Then from theorem 75, we know that for any $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y}-\mathbf{x}) - \frac{1}{2}c||\mathbf{y}+\mathbf{x}||^2$$
$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla^T f(\mathbf{y})(\mathbf{x}-\mathbf{y}) - \frac{1}{2}c||\mathbf{x}+\mathbf{y}||^2$$

Adding the two inequalities, we get (4.55). If f is convex, the inequalities hold with $c = 0$, yielding (4.54). If f is strictly convex, the inequalities will be strict, yielding (4.54).

Sufficiency: Suppose ∇f is monotone. For any fixed $\mathbf{x}, \mathbf{y} \in \mathcal{D}$, consider the function $\phi(t) = f(\mathbf{x} + t(\mathbf{y}-\mathbf{x}))$. By the mean value theorem applied to $\phi(t)$, we should have for some $t \in (0,1)$,

$$\phi(1) - \phi(0) = \phi'(t) \tag{4.56}$$

Letting $\mathbf{z} = \mathbf{x} + t(\mathbf{y}-\mathbf{x})$, (4.56) translates to

$$f(\mathbf{y}) - f(\mathbf{x}) = \nabla^T f(\mathbf{z})(\mathbf{y}-\mathbf{x}) \tag{4.57}$$

Also, by definition of monotonicity of ∇f, (from (4.53)),

$$(\nabla f(\mathbf{z}) - \nabla f(\mathbf{x}))^T(\mathbf{y}-\mathbf{x}) = \frac{1}{t}(\nabla f(\mathbf{z}) - \nabla f(\mathbf{x}))^T(\mathbf{z}-\mathbf{x}) \geq 0 \tag{4.58}$$

Combining (4.57) with (4.58), we get,

$$\begin{aligned} f(\mathbf{y}) - f(\mathbf{x}) &= (\nabla f(\mathbf{z}) - f(\mathbf{x}))^T(\mathbf{y}-\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y}-\mathbf{x}) \\ &\geq \nabla^T f(\mathbf{x})(\mathbf{y}-\mathbf{x}) \end{aligned} \tag{4.59}$$

By theorem 75, this inequality proves that f is convex. Strict convexity can be similarly proved by using the strict inequality in (4.58) inherited from strict monotonicity, and letting the strict inequality follow through to (4.59). For the case of strong convexity, from (4.55), we have

$$\begin{aligned} \phi'(t) - \phi'(0) &= (\nabla f(\mathbf{z}) - f(\mathbf{x}))^T(\mathbf{y}-\mathbf{x}) \\ = \frac{1}{t}(\nabla f(\mathbf{z}) - f(\mathbf{x}))^T(\mathbf{z}-\mathbf{x}) &\geq \frac{1}{t}c||\mathbf{z}-\mathbf{x}||^2 = ct||\mathbf{y}-\mathbf{x}||^2 \end{aligned} \tag{4.60}$$

$$\phi(1) - \phi(0) - \phi'(0) = \int_0^1 [\phi'(t) - \phi'(0)]dt \geq \frac{1}{2}c||\mathbf{y}-\mathbf{x}||^2 \tag{4.61}$$

which translates to

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{1}{2}c||\mathbf{y}-\mathbf{x}||^2$$

Lagrange dual function

$\max_{\lambda \succeq 0, \nu} g(\lambda,\nu) = \min_{\lambda \succeq 0, \nu} -g(\lambda,\nu)$

convex function in λ & ν

$\therefore$ Dual optimization prob is always a concave max / convex minimization problem

Lagrange dual function: $g : \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$,

$$\begin{aligned} g(\lambda, \nu) &= \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) \\ &= \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x) \right) = \Omega\left(f_0(x) + \sum \lambda_i f_i(x) + \sum \nu_i h_i(x)\right) \end{aligned}$$

H/W: Prove

g is concave, can be $-\infty$ for some λ, ν

lower bound property: if $\lambda \succeq 0$, then $g(\lambda, \nu) \leq p^\star$

($\lambda \succeq 0$ by default means $\lambda \in \mathbb{R}^n_+$)

proof: if $\tilde{x}$ is feasible and $\lambda \succeq 0$, then

$$f_0(\tilde{x}) \geq L(\tilde{x}, \lambda, \nu) \geq \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) = g(\lambda, \nu)$$

minimizing over all feasible $\tilde{x}$ gives $p^\star \geq g(\lambda, \nu)$

Least-norm solution of linear equations

$$\begin{array}{ll} \text{minimize} & x^T x \\ \text{subject to} & Ax = b \end{array}$$

dual function

- Lagrangian is $L(x, \nu) = x^T x + \nu^T (Ax - b)$
- to minimize L over x, set gradient equal to zero:

Using necessary condition.

$$\nabla_x L(x, \nu) = 2x + A^T \nu = 0 \quad \Longrightarrow \quad x = -(1/2) A^T \nu$$

- plug in in L to obtain g:

$$g(\nu) = L((-1/2)A^T \nu, \nu) = -\frac{1}{4} \nu^T A A^T \nu - b^T \nu$$

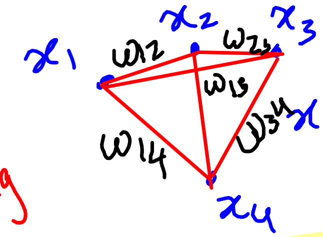
a concave function of ν

Quadratic in ν

lower bound property: $p^\star \geq -(1/4)\nu^T A A^T \nu - b^T \nu$ for all ν

$$W = \begin{bmatrix} w_{11} & w_{12} & \\ w_{21} & \ddots & \\ & & w_{44} \end{bmatrix}$$

Two-way partitioning



$x_i \in \{-1, +1\}$

Ising Models

$$\begin{array}{ll} \text{minimize} & x^TWx \quad \rightarrow \text{energy level of config} \\ \text{subject to} & \underbrace{x_i^2 = 1}_{\text{Non convex domain}}, \quad i = 1, \ldots, n \end{array}$$

- a nonconvex problem; feasible set contains 2^n discrete points
- interpretation: partition $\{1, \ldots, n\}$ in two sets; W_{ij} is cost of assigning i, j to the same set; $-W_{ij}$ is cost of assigning to different sets

dual function

$$\begin{aligned} g(\nu) = \inf_x (x^TWx + \sum_i \nu_i(x_i^2 - 1)) &= \inf_x x^T(W + \mathbf{diag}(\nu))x - \mathbf{1}^T\nu \\ &= \begin{cases} -\mathbf{1}^T\nu & W + \mathbf{diag}(\nu) \succeq 0 \\ -\infty & \text{otherwise} \end{cases} \end{aligned}$$

p.s.d constraint

lower bound property: $p^\star \geq -\mathbf{1}^T\nu$ if $W + \mathbf{diag}(\nu) \succeq 0$

example: $\nu = -\lambda_{\min}(W)\mathbf{1}$ gives bound $p^\star \geq n\lambda_{\min}(W)$

$W + \begin{bmatrix} -\lambda_{min} & 0 \\ 0 & -\lambda_{min} \end{bmatrix} \succeq 0$ substitute

Lagrange dual and conjugate function

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & Ax \preceq b, \quad Cx = d \end{array}$$

dual function

$$\begin{aligned} g(\lambda, \nu) &= \inf_{x \in \mathbf{dom}\, f_0} \left(f_0(x) + (A^T\lambda + C^T\nu)^Tx - b^T\lambda - d^T\nu \right) \\ &= -f_0^*(-A^T\lambda - C^T\nu) - b^T\lambda - d^T\nu \end{aligned}$$

- recall definition of conjugate $f^*(y) = \sup_{x \in \mathbf{dom}\, f}(y^Tx - f(x))$
- simplifies derivation of dual if conjugate of f_0 is kown

example: entropy maximization

$$f_0(x) = \sum_{i=1}^n x_i \log x_i, \qquad f_0^*(y) = \sum_{i=1}^n e^{y_i - 1}$$

Reminds you of Entropy classifier Logistic regression

The dual problem

Lagrange dual problem

$$\begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \succeq 0 \end{array}$$

- finds best lower bound on $p^\star$, obtained from Lagrange dual function
- a convex optimization problem; optimal value denoted $d^\star$
- λ, ν are dual feasible if $\lambda \succeq 0$, $(\lambda, \nu) \in \mathbf{dom}\, g$
- often simplified by making implicit constraint $(\lambda, \nu) \in \mathbf{dom}\, g$ explicit

example: standard form LP and its dual (page 5–5) (also seen for conic Prog CP)

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b \\ & x \succeq 0 \end{array} \quad \equiv \quad \begin{array}{ll} \text{maximize} & -b^T \nu \\ \text{subject to} & A^T \nu + c \succeq 0 \end{array}$$

Recap: If both primal & dual are feasible & one of them is strictly feasible ⇒ zero duality (strong duality)

Weak and strong duality

weak duality: $d^\star \leq p^\star$

- always holds (for convex and nonconvex problems)
- can be used to find nontrivial lower bounds for difficult problems

  for example, solving the SDP

$$\begin{array}{ll} \text{maximize} & -\mathbf{1}^T \nu \\ \text{subject to} & W + \mathbf{diag}(\nu) \succeq 0 \end{array}$$

  gives a lower bound for the two-way partitioning problem on page 5–7

strong duality: $d^\star = p^\star$

- does not hold in general
- (usually) holds for convex problems
- conditions that guarantee strong duality in convex problems are called **constraint qualifications**

Slater's constraint qualification

strong duality holds for a convex problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \le 0, \quad i = 1, \dots, m \\ & Ax = b \end{array}$$

Akin to (strong) conic duality

if it is strictly feasible, *i.e.*,

$$\exists x \in \mathbf{int}\, \mathcal{D}: \qquad f_i(x) < 0, \quad i = 1, \dots, m, \qquad Ax = b$$

- also guarantees that the dual optimum is attained (if $p^\star > -\infty$)
- can be sharpened: *e.g.*, can replace $\mathbf{int}\, \mathcal{D}$ with $\mathbf{relint}\, \mathcal{D}$ (interior relative to affine hull); linear inequalities do not need to hold with strict inequality, . . .
- there exist many other types of constraint qualifications

Inequality form LP

primal problem

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \preceq b \end{array}$$

dual function

$$g(\lambda) = \inf_x \left((c + A^T\lambda)^T x - b^T \lambda \right) = \begin{cases} -b^T\lambda & A^T\lambda + c = 0 \\ -\infty & \text{otherwise} \end{cases}$$

dual problem

$$\begin{array}{ll} \text{maximize} & -b^T \lambda \\ \text{subject to} & A^T\lambda + c = 0, \quad \lambda \succeq 0 \end{array}$$

- from Slater's condition: $p^\star = d^\star$ if $A\tilde{x} \prec b$ for some $\tilde{x}$
- in fact, $p^\star = d^\star$ except when primal and dual are infeasible

Complementary slackness

assume strong duality holds, $x^\star$ is primal optimal, $(\lambda^\star, \nu^\star)$ is dual optimal

$$\begin{aligned} f_0(x^\star) = g(\lambda^\star, \nu^\star) &= \inf_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^\star f_i(x) + \sum_{i=1}^p \nu_i^\star h_i(x) \right) \\ &\leq f_0(x^\star) + \underbrace{\sum_{i=1}^m \lambda_i^\star f_i(x^\star)}_{} + \underbrace{\sum_{i=1}^p \nu_i^\star h_i(x^\star)}_{h_j(x^\star)=0} \\ &\leq f_0(x^\star) \end{aligned}$$

$\lambda_i^\star \geq 0$ $f_i(x^\star) \leq 0$

$\because \lambda_i^\star$ is dual feasible $\because x^\star$ is primal feasible

hence, the two inequalities hold with equality

- $x^\star$ minimizes $L(x, \lambda^\star, \nu^\star)$
- $\lambda_i^\star f_i(x^\star) = 0$ for $i = 1, \ldots, m$ (known as complementary slackness):

$$\lambda_i^\star > 0 \Longrightarrow f_i(x^\star) = 0, \qquad f_i(x^\star) < 0 \Longrightarrow \lambda_i^\star = 0$$

Karush-Kuhn-Tucker (KKT) conditions

the following four conditions are called KKT conditions (for a problem with differentiable f_i, h_i):

1. primal constraints: $f_i(x) \leq 0$, $i = 1, \ldots, m$, $h_i(x) = 0$, $i = 1, \ldots, p$
2. dual constraints: $\lambda \succeq 0$
3. complementary slackness: $\lambda_i f_i(x) = 0$, $i = 1, \ldots, m$
4. gradient of Lagrangian with respect to x vanishes:

$$\nabla f_0(x) + \sum_{i=1}^m \lambda_i \nabla f_i(x) + \sum_{i=1}^p \nu_i \nabla h_i(x) = 0$$

from page 5–17: if strong duality holds and x, λ, ν are optimal, then they must satisfy the KKT conditions

KKT conditions for convex problem

if $\tilde{x}$, $\tilde{\lambda}$, $\tilde{\nu}$ satisfy KKT for a convex problem, then they are optimal:

- from complementary slackness: $f_0(\tilde{x}) = L(\tilde{x}, \tilde{\lambda}, \tilde{\nu})$
- from 4th condition (and convexity): $g(\tilde{\lambda}, \tilde{\nu}) = L(\tilde{x}, \tilde{\lambda}, \tilde{\nu})$

hence, $f_0(\tilde{x}) = g(\tilde{\lambda}, \tilde{\nu})$

if **Slater's condition** is satisfied: "only if" assumes differentiability of f, g_i & h_j

x is optimal if and only if there exist λ, ν that satisfy KKT conditions

- recall that Slater implies strong duality, and dual optimum is attained
- generalizes optimality condition $\nabla f_0(x) = 0$ for unconstrained problem

example: water-filling (assume $\alpha_i > 0$)

$$\begin{array}{ll} \text{minimize} & -\sum_{i=1}^n \log(x_i + \alpha_i) \\ \text{subject to} & x \succeq 0, \quad \mathbf{1}^T x = 1 \end{array}$$

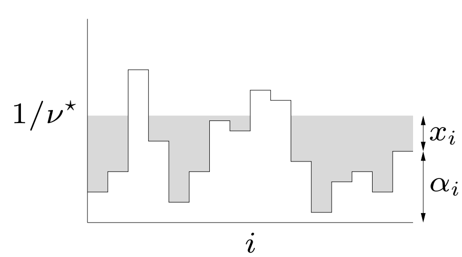
x is optimal iff $x \succeq 0$, $\mathbf{1}^T x = 1$, and there exist $\lambda \in \mathbf{R}^n$, $\nu \in \mathbf{R}$ such that

$$\lambda \succeq 0, \qquad \lambda_i x_i = 0, \qquad \frac{1}{x_i + \alpha_i} + \lambda_i = \nu$$

- if $\nu < 1/\alpha_i$: $\lambda_i = 0$ and $x_i = 1/\nu - \alpha_i$
- if $\nu \geq 1/\alpha_i$: $\lambda_i = \nu - 1/\alpha_i$ and $x_i = 0$
- determine ν from $\mathbf{1}^T x = \sum_{i=1}^n \max\{0, 1/\nu - \alpha_i\} = 1$

interpretation

- n patches; level of patch i is at height α_i
- flood area with unit amount of water
- resulting level is $1/\nu^\star$



Algorithms inspired by Lagrange duality

$$\boxed{\begin{aligned} &\min \ f(x) \\ &\text{s.t } Ax = b \end{aligned}}$$

KKT: $L(x,\nu) = f(x) + \nu^T(Ax-b)$

$\nabla_x L(x^*,\nu^*) = \nabla f(x^*) + A^T\nu^* = 0$ $\{*\}$

If $Ax \leq b$ then additionally: $\nu^* \geq 0$

& :. $\nu^{(k+1)} = [\nu^{(k)} + t^{(k)}(Ax^{(k+1)} - b)]_+$

componentwise thresholding

Note: $\partial_\nu L(x,\nu) = Ax - b$

Dual ascent

$$\boxed{\begin{aligned} x^{(k+1)} &= \underset{x}{\operatorname{argmin}}\ L(x,\nu^{(k)}) = \underset{x}{\operatorname{argmin}}\ f(x) + (\nu^{(k)})^T Ax \\ \nu^{(k+1)} &= \nu^{(k)} + t^{(k)}\partial_\nu L(x^{(k+1)},\nu^{(k)}) = \nu^{(k)} + t^{(k)}(Ax^{(k+1)} - b) \end{aligned}}$$

If strong duality holds & $\nu^{(k)} \to \nu^{(*)}$ then

$x^{(*)} = \underset{x}{\operatorname{argmin}}\ L(x,\nu^*)$ provided minimizer is unique

$O(1/k)$ convergence if f is Lipschitz continuous with const d & provided $t^{(k)} \leq d \ \forall k$.

Dual decomposition

$$\boxed{f(x) = \sum_{i=1}^{b} f_i(x_i)}$$

$x = [x_1, x_2 \ldots x_b]$ .. b blocks of variables

Simplification of dual ascent (useful for parallelization)

More robust: Method of Augmented Lagrangian

Apply Dual descent to

$$\min_x \ f(x) + \frac{\rho}{2}\|Ax-b\|^2$$
$$\text{s.t } Ax = b$$

$$x^{(k+1)} = \underset{x}{\operatorname{argmin}}\ \underbrace{f(x) + \frac{\rho}{2}\|Ax-b\|^2 + (\nu^{(k)})^T(Ax-b)}_{L_\rho(x,\nu)}$$

$$\nu^{(k+1)} = \nu^k + \rho(Ax^{(k)} - b)$$

$x^{(k+1)} = \arg\min_{x} f(x) + \nu^{(k)T} A x$

$= \arg\min_{x_1 .. x_b} \sum_i f_i(x_i) + \nu^{(k)T} \left(\sum_i A_i x_i\right)$

$(A = [A_1 \; A_2 \ldots A_b])$

Thus:

$$\boxed{x_i^{(k+1)} = \arg\min_{x_i} f_i(x_i) + \nu^{(k)T} A_i x_i}$$

for each block, possibly in parallel after broadcasting $\nu^{(k)}$

$$\boxed{\nu^{(k+1)} = \nu^{(k)} + t^{(k)}(Ax^{(k+1)} - b)}$$

After gathering $x_i^{(k+1)}$ from all nodes & updating $\nu^{(k+1)}$ for again broadcasting

$t^{(k)} = \rho$ motivated by observation that necessary condition for $x^{(k+1)}$ is

$$0 \in \partial f(x^{(k+1)}) + A^T \big(\underbrace{\nu^{(k)} + \rho(Ax^{(k)} - b)}_{y^{(k+1)}}\big)$$

i.e. $0 \in \partial f(x^{(k+1)}) + A^T y^{(k+1)}$

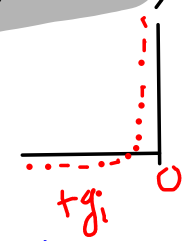
Comparing with {*} expression for $y^{(k+1)}$ can be update rule for $\nu^{(k+1)}$!

Dual descent on Augmented Lagrangian has better conver-gence rate

Interior Point Methods (IPM)

(Barrier method... like augmented Lagrangian)

$$x^{(k+1)} = \arg\min_{x} f(x) - \frac{1}{t^{(k)}} \sum \log(-g_i(x^k))$$



ADMM (Alternating direction Method of Multipliers)

$f(x) = \sum f_i(x_i)$, $L_\rho(x_1 .. x_b, \nu) = \sum f_i(x_i) + \frac{\rho}{2}\|Ax - b\|^2 + \nu^T(Ax - b)$

$x_1^{(k+1)} = \arg\min_{x_1} L_\rho(x_1, x_2^k, \ldots \nu^k)$, $x_2^{(k+1)} = \arg\min_{x_2} L_\rho(x_1^{(k+1)}, x_2, x_3^{(k)} \ldots \nu^k)$

$x_3^{(k+1)} = \arg\min_{x_3} L_\rho(x_1^{(k+1)}, x_2^{(k+1)}, x_3, x_4^{(k)} \ldots \nu^{(k)}) \ldots x_b^{(k+1)} = \arg\min_{x_b} L_\rho(x_1^{(k+1)}, \ldots x_{b-1}^{(k+1)}, x_b, \nu^{(k)})$

$\nu^{(k+1)} = \nu^k + \rho(A_1 x_1^{(k+1)} + A_2 x_2^{(k+1)} \ldots + A_b x_b^{(k+1)} - b)$

Other methods (First order: Require only upto $\nabla f(x)$)

① Mirror descent

Recap: Gradient descent (subgradient descent)

$$x^{(k+1)} = \underset{x}{\operatorname{argmin}}\ f(x^k) + \partial^T f(x^k)(x - x^k) + \frac{t}{2}\|x - x^k\|^2$$

Mirror descent:

$$x^{(k+1)} = \underset{x}{\operatorname{argmin}}\ f(x^k) + \partial^T f(x^k)(x - x^k) + \Delta_g(x, x^k)$$

(Bregman Divergence)

② Accelerated gradient descent methods

$$x^{(k+1)} = F(x^{(k)}, x^{(k-1)}, \ldots)$$

- First iteration is just usual proximal gradient descent update
    - After that, updates carry some "momentum" from previous iterations
    - Example: Conjugate gradient methods: Subtracts previous descent direction from the current gradient to give a current descent direction that is steep but is nearly orthogonal to previous descent directions. Pages 317 to 324 of http://www.cse.iitb.ac.in/~cs709/notes/BasicsOfConvexOptimization.pdf. In particular, see Figure 4.55
    - LBFGS | BFGS (see notes ...)

③ ACTIVE SET METHOD

http://www.cse.iitb.ac.in/~cs709/notes/quadraticOpt-PrimalActiveSet.pdf

④ CUTTING PLANE METHOD

http://www.cse.iitb.ac.in/~cs709/notes/kellysCuttingPlaneAlgo.pdf

5) Second order methods (Newton's method)

http://www.cse.iitb.ac.in/~cs709/notes/BasicsOfConvexOptimization.pdf (pg 305)