LECTURE 24

LECTURE OUTLINE

- Extensions of proximal and projection ideas
- Nonquadratic proximal algorithms
- Entropy minimization algorithm
- Exponential augmented Lagrangian method
- Entropic descent algorithm

****

References:

- On-line chapter on algorithms

- Bertsekas, D. P., 1999. Nonlinear Programming, Athena Scientific, Belmont, MA.

- Beck, A., and Teboulle, M., 2003. "Mirror Descent and Nonlinear Projected Subgradient Methods for Convex Optimization," Operations Research Letters, Vol. 31, pp. 167-175.

GENERALIZED PROXIMAL-RELATED ALGS

- Introduce a general regularization term D_k:

$$x_{k+1} \in \arg\min_{x \in X} \Big\{ f(x) + D_k(x, x_k) \Big\}$$

- All the ideas extend to the nonquadratic case (although the analysis may not be trivial).
- In particular we have generalizations as follows:
  - Dual proximal algorithms (based on Fenchel duality)
  - Augmented Lagrangian methods with nonquadratic penalty functions
  - Combinations with polyhedral approximations (bundle-type methods)
  - Proximal gradient method
  - Incremental subgradient-proximal methods
  - Gradient projection algorithms with "nonquadratic metric"
- We may look also at what happens when f is not convex.

SPECIAL CASE: ENTROPY REGULARIZATION

$$D_k(x,y) = \begin{cases} \frac{1}{c_k} \sum_{i=1}^n x^i \left(\ln\left(\frac{x^i}{y^i}\right) - 1 \right) & \text{if } x > 0,\ y > 0, \\ \infty & \text{otherwise} \end{cases}$$

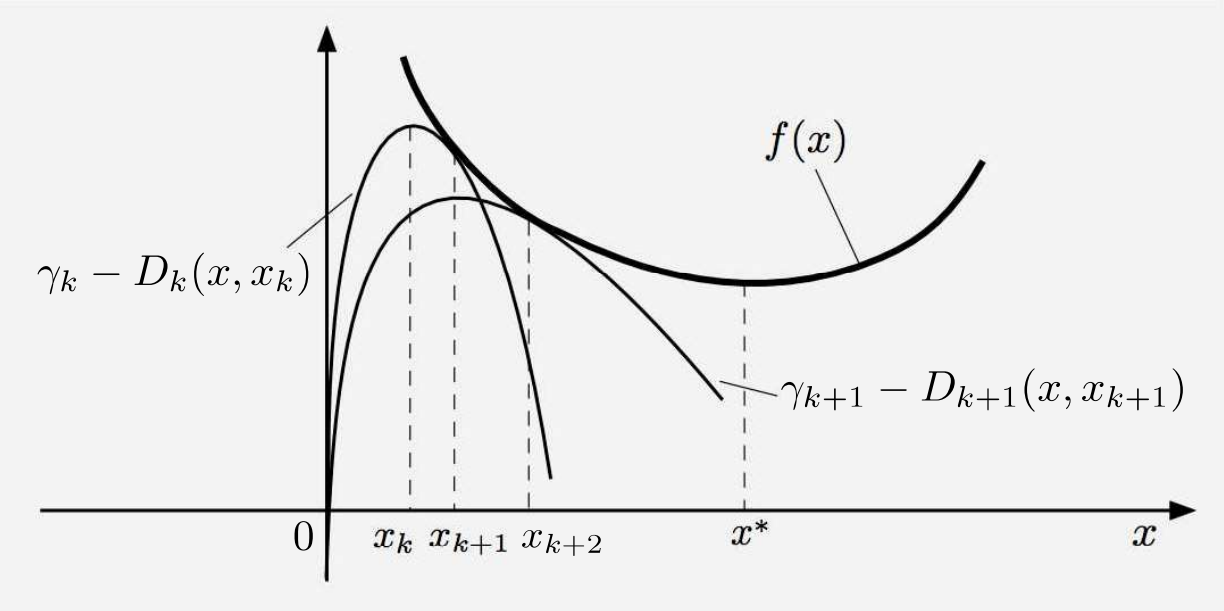
- Also written as

$$D_k(x,y) = \frac{1}{c_k} \sum_{i=1}^n y^i \phi_i \left(\frac{x^i}{y^i} \right),$$

where

$$\phi(x) = \begin{cases} x\big(\ln(x) - 1\big) & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ \infty & \text{if } x < 0. \end{cases}$$

- Proximal algorithm:

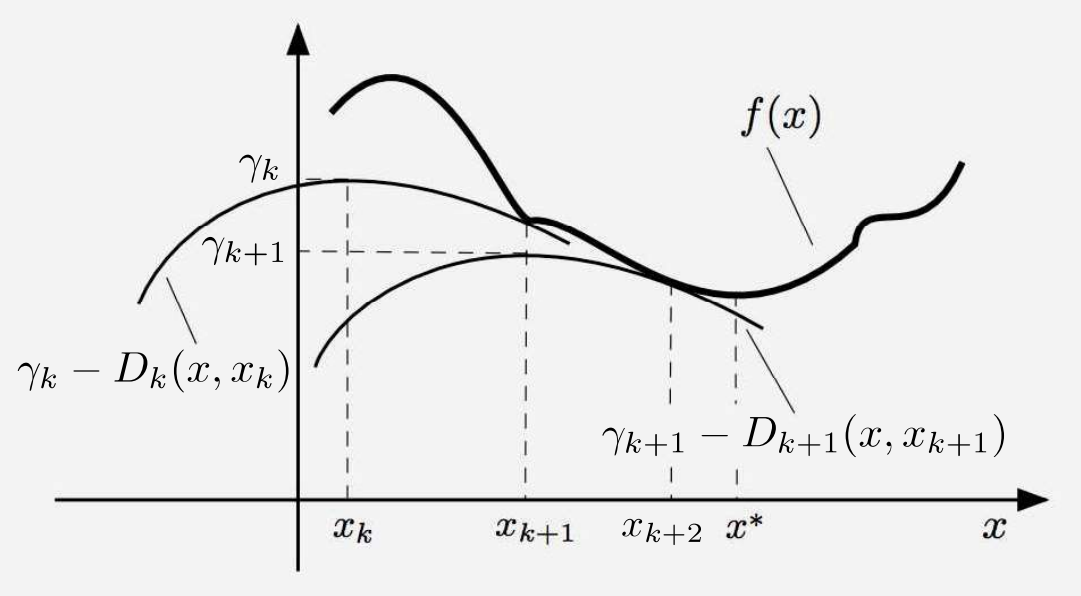


GENERALIZED PROXIMAL ALGORITHM

- Introduce a general regularization term D_k : $\Re^{2n} \mapsto (-\infty, \infty]$:

$$x_{k+1} \in \arg\min_{x \in \Re^n} \Big\{ f(x) + D_k(x, x_k) \Big\}$$

- Consider a general cost function f



- Assume attainment of min (but this is not automatically guaranteed)

- Complex/unreliable behavior when f is nonconvex

SOME GUARANTEES ON GOOD BEHAVIOR

- Assume "stabilization property"

$$D_k(x, x_k) \geq D_k(x_k, x_k), \qquad \forall\ x \in \Re^n,\ k \tag{1}$$

Then we have a cost improvement property:

$$\begin{aligned} f(x_{k+1}) &\leq f(x_{k+1}) + D_k(x_{k+1}, x_k) - D_k(x_k, x_k) \\ &\leq f(x_k) + D_k(x_k, x_k) - D_k(x_k, x_k) \\ &= f(x_k) \end{aligned} \tag{2}$$

- Assume algorithm stops only when x_k is in optimal solution set X^*, i.e.,

$$x_k \in \arg\min_{x \in \Re^n} \big\{ f(x) + D_k(x, x_k) \big\} \quad \Rightarrow \quad x_k \in X^*$$

- Then strict cost improvement for $x_k \notin X^*$ [the second inequality in (2) is strict].

- Guaranteed if f is convex and:

  (a) $D_k(\cdot, x_k)$ satisfies (1), and is convex and differentiable at x_k.

  (b) $\text{ri}\big(\text{dom}(f)\big) \cap \text{ri}\big(\text{dom}(D_k(\cdot, x_k))\big) \neq \emptyset$.

EXAMPLES

- Bregman distance function

$$D_k(x,y) = \frac{1}{c_k}\big(\phi(x) - \phi(y) - \nabla\phi(y)'(x-y)\big),$$

where $\phi : \Re^n \mapsto (-\infty, \infty]$ is a convex function, differentiable within an open set containing $\mathrm{dom}(f)$, and c_k is a positive penalty parameter. Special cases: quadratic and entropy functions.

- Majorization-Minimization algorithm:

$$D_k(x,y) = M_k(x,y) - M_k(y,y),$$

where M satisfies

$$M_k(y,y) = f(y), \qquad \forall\ y \in \Re^n,\ k = 0, 1,$$

$$M_k(x,x_k) \geq f(x_k), \qquad \forall\ x \in \Re^n,\ k = 0, 1, \ldots$$

- Example for case $f(x) = R(x) + \|Ax - b\|^2$, where R is a convex regularization function

$$M(x,y) = R(x) + \|Ax - b\|^2 - \|Ax - Ay\|^2 + \|x - y\|^2$$

- Expectation-Maximization (EM) algorithm (special context in inference, f nonconvex)

DUAL PROXIMAL MINIMIZATION

- The proximal iteration can be written in the Fenchel form: $\min_x\{f_1(x)+f_2(x)\}$ with

$$f_1(x)=f(x), \qquad f_2(x)=D_k(x;x_k)$$

- The Fenchel dual is

$$\begin{aligned} &\text{minimize} \quad f^\star(\lambda)+D_k^\star(\lambda;x_k)\\ &\text{subject to} \quad \lambda\in\Re^n \end{aligned}$$

where $D_k^\star(\cdot;x_k)$ is the conjugate of $D_k(\cdot;x_k)$:

$$D_k^\star(\lambda;x_k)=\sup_{x\in\Re^n}\Big\{-\lambda'x-D_k(x;x_k)\Big\}$$

- If $D_k(\cdot;x_k)$ or $D_k^\star(\cdot;x_k)$ is real-valued, there is no duality gap.
- Can use the Fenchel dual for a dual proximal implementation.

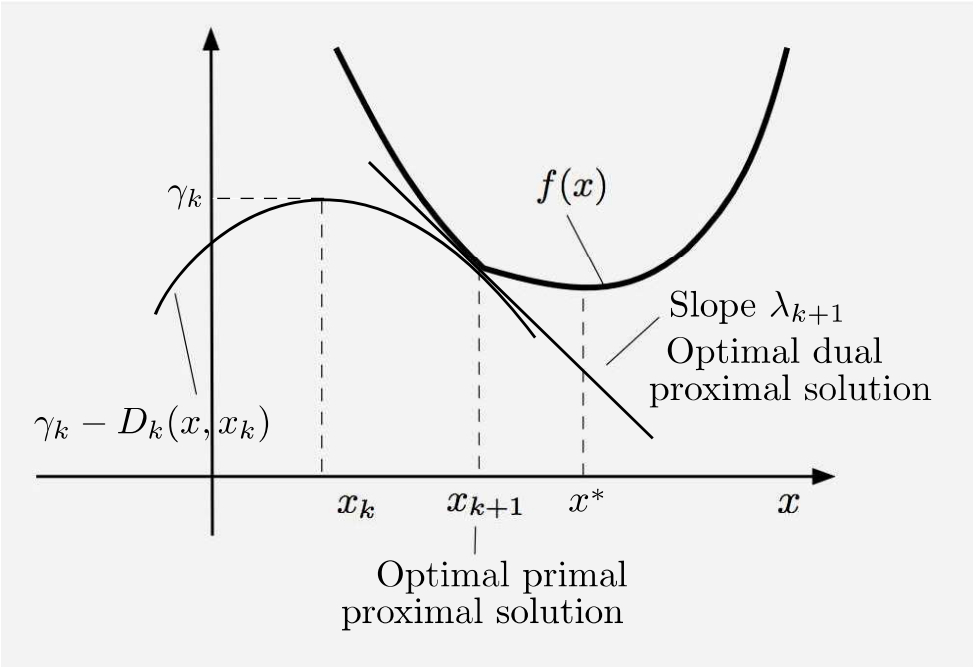
DUAL IMPLEMENTATION

- We can solve the Fenchel-dual problem instead of the primal at each iteration:

$$\lambda_{k+1} = \arg \min_{\lambda \in \Re^n} \{f^\star(\lambda) + D_k^\star(\lambda; x_k)\}$$

- Primal-dual optimal pair (x_{k+1}, λ_{k+1}) are related by the "differentiation" condition:

$$\lambda_{k+1} \in \partial D_k(x_{k+1}; x_k) \quad \text{or} \quad x_{k+1} \in \partial D_k^\star(\lambda_{k+1}; x_k)$$



- The primal and dual algorithms generate identical sequences $\{x_k, \lambda_k\}$.

- Special cases: Augmented Lagrangian methods with nonquadratic penalty functions.

ENTROPY/EXPONENTIAL DUALITY

- A special case involving entropy regularization:

$$x_{k+1} \in \arg\min_{x \in X} \left\{ f(x) + \frac{1}{c_k} \sum_{i=1}^{n} x^i \left(\ln\left(\frac{x^i}{x_k^i} \right) - 1 \right) \right\}$$

where $x_k > 0$.

- Fenchel duality $\Rightarrow$ Augmented Lagrangian method
- Note: The conjugate of the logarithmic

$$h(x) = \begin{cases} x\big(\ln(x) - 1\big) & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ \infty & \text{if } x < 0, \end{cases}$$

is the exponential $h^\star(y) = e^y$.

- The dual (augmented Lagrangian) problem is

$$u_{k+1} \in \arg\min_{u \in \Re^n} \left\{ f^\star(u) + \frac{1}{c_k} \sum_{i=1}^{n} x_k^i e^{c_k u^i} \right\}$$

The proximal/multiplier iteration is

$$x_{k+1}^i = x_k^i e^{c_k u_{k+1}^i}, \qquad i = 1, \ldots, n$$

EXPONENTIAL AUGMENTED LAGRANGIAN

- A special case for the convex problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad g_1(x) \leq 0, \ldots, g_r(x) \leq 0, \quad x \in X \end{aligned}$$

- Apply proximal to the (Langrange) dual problem. It consists of unconstrained minimizations

$$x_k \in \arg\min_{x \in X} \left\{ f(x) + \frac{1}{c_k} \sum_{j=1}^{r} \mu_k^j e^{c_k g_j(x)} \right\},$$

followed by the multiplier iterations

$$\mu_{k+1}^j = \mu_k^j e^{c_k^j g_j(x_k)}, \qquad j = 1, \ldots, r$$

- Note: We must have $\mu_0 > 0$, which implies $\mu_k > 0$ for all k.
- Theoretical convergence properties are similar to the quadratic augmented Lagrangian method.
- The exponential is twice differentiable, hence more suitable for Newton-like methods.

NONLINEAR PROJECTION ALGORITHM

- Subgradient projection with general regularization D_k:

$$x_{k+1} \in \arg\min_{x \in X} \left\{ f(x_k) + \tilde{\nabla} f(x_k)'(x - x_k) + D_k(x, x_k) \right\}$$

where $\tilde{\nabla} f(x_k)$ is a subgradient of f at x_k. Also called mirror descent method.

- Linearization of f simplifies the minimization.

- The use of nonquadratic linearization is useful in problems with special structure.

- Entropic descent method: Minimize $f(x)$ over the unit simplex $X = \left\{ x \geq 0 \mid \sum_{i=1}^n x^i = 1 \right\}$.

- Method:

$$x_{k+1} \in \arg\min_{x \in X} \sum_{i=1}^n \left(x^i \tilde{\nabla}_i f(x_k) + \frac{1}{\alpha_k} x^i \left(\ln \left(\frac{x^i}{x_k^i} \right) - 1 \right) \right)$$

where $\tilde{\nabla}_i f(x_k)$ are the components of $\tilde{\nabla} f(x_k)$.

- This minimization can be done in closed form:

$$x_{k+1}^i = \frac{x_k^i e^{-\alpha_k \tilde{\nabla}_i f(x_k)}}{\sum_{j=1}^n x_k^j e^{-\alpha_k \tilde{\nabla}_j f(x_k)}}, \qquad i = 1, \ldots, n$$