

5. Gradient projection

- projected gradient
- examples
- convergence analysis
- dual gradient methods

5-1

(Sub-)gradient projection

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in C \end{array}$$

f convex with $\text{dom } f = \mathbf{R}^n$

C a closed convex set; we denote by P_C (or just P) the projection on C

- **gradient projection** (f differentiable)

$$x^+ = P(x - t\nabla f(x))$$

can also be included in Nesterov's fast gradient method

- **subgradient projection** (f nondifferentiable)

$$x^+ = P(x - tg) \quad g \in \partial f(x)$$

interesting as large-scale algorithms if the projection is inexpensive

Affine sets

hyperplane: $C = \{x \mid a^T x = b\}$ (with $a \neq 0$)

$$P(x) = x + \frac{b - a^T x}{\|a\|_2^2} a$$

affine set: $C = \{x \mid Ax = b\}$ (with $A \in \mathbf{R}^{p \times n}$ and $\text{rank}(A) = p$)

$$P(x) = x + A^T(AA^T)^{-1}(b - Ax)$$

inexpensive if $p \ll n$, or $AA^T = I, \dots$

Simple polyhedral sets

halfspace: $C = \{x \mid a^T x \leq b\}$ (with $a \neq 0$)

$$P(x) = x + \frac{b - a^T x}{\|a\|_2^2} a \quad (\text{if } a^T x > b), \quad P(x) = x \quad (\text{otherwise})$$

rectangle: $C = \{x \mid l \preceq x \preceq u\}$

$$P(x)_i = \begin{cases} l_i & x_i \leq l_i \\ x_i & l_i \leq x_i \leq u_i \\ u_i & x_i \geq u_i \end{cases}$$

nonnegative orthant: $C = \mathbf{R}_+^n$

$$P(x) = x_+ \quad (x_+ \text{ is componentwise max of } 0 \text{ and } x)$$

probability simplex: $C = \{x \mid \mathbf{1}^T x = 1, x \succeq 0\}$

$$P(x) = (x - \lambda \mathbf{1})_+$$

where λ is the solution of the equation

$$\mathbf{1}^T (x - \lambda \mathbf{1})_+ = \sum_{i=1}^n \max\{0, x_k - \lambda\} = 1$$

intersection of hyperplane and rectangle: $C = \{x \mid a^T x = b, l \preceq x \preceq u\}$

$$P(x) = P_{[l,u]}(x - \lambda a)$$

where λ is the solution of

$$a^T P_{[l,u]}(x - \lambda a) = b$$

Norm balls

Euclidean ball: $C = \{x \mid \|x\|_2 \leq 1\}$

$$P(x) = \frac{1}{\|x\|_2} x \quad (\text{if } \|x\|_2 \geq 1), \quad P(x) = x \quad (\text{otherwise})$$

1-norm ball: $C = \{x \mid \|x\|_1 \leq 1\}$

$$P(x)_k = \begin{cases} x_k - \lambda & x_k > \lambda \\ 0 & -\lambda \leq x_k \leq \lambda \\ x_k + \lambda & x_k < -\lambda \end{cases}$$

$\lambda = 0$ if $\|x\|_1 \leq 1$; otherwise λ is the solution of the equation

$$\sum_{k=1}^n \max\{|x_k| - \lambda, 0\} = 1$$

Simple cones

second order cone $C = \{(x, t) \in \mathbf{R}^{n \times 1} \mid \|x\|_2 \leq t\}$

$$P(x, t) = (x, t) \quad (\text{if } \|x\|_2 \leq t), \quad P(x, t) = (0, 0) \quad (\text{if } \|x\|_2 \leq -t)$$

and

$$P(x, t) = \frac{t + \|x\|_2}{2} \begin{bmatrix} x/\|x\|_2 \\ 1 \end{bmatrix} \quad (\text{if } -t < \|x\|_2 < t)$$

positive semidefinite cone $C = \mathbf{S}_+^n$

$$P(X) = \sum_{i=1}^n \max\{0, \lambda_i\} q_i q_i^T$$

if $X = \sum_{i=1}^n \lambda_i q_i q_i^T$ is the eigenvalue decomposition of X

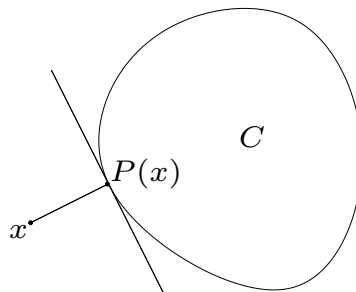
Gradient projection

5-7

Euclidean projection

Euclidean projection on closed convex set C

$$P(x) = \operatorname{argmin}_{z \in C} \|z - x\|_2^2$$



- $P(x)$ exists and is unique for all x
- optimality condition:

$$(x - P(x))^T (z - P(x)) \leq 0 \quad \forall z \in C$$

Gradient projection

5-8

Gradient map

definition (for f convex, differentiable, with $\text{dom } f = \mathbf{R}^n$)

$$G_t(x) = \frac{1}{t}(x - P(x - t\nabla f(x)))$$

properties

- allows us to write the projection of $x - t\nabla f(x)$ as

$$P(x - t\nabla f(x)) = x - tG_t(x)$$

- from optimality condition on page 5–8

$$(G_t(x) - \nabla f(x))^T (z - x + tG_t(x)) \leq 0 \quad \forall z \in C$$

Optimality condition

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in C \end{array}$$

optimality condition: x is optimal if and only if

$$G_t(x) = 0$$

(for any $t > 0$)

proof: from previous page, $G_t(x) = 0$ is equivalent to

$$x \in C, \quad \nabla f(x)^T (z - x) \geq 0 \quad \forall z \in C$$

Projected gradient methods

gradient method

choose $x^{(0)} \in C$ and repeat

$$x^{(k)} = P\left(x^{(k-1)} - t_k \nabla f(x^{(k-1)})\right), \quad k = 1, 2, \dots$$

Nesterov's gradient method

choose $x^{(0)} \in C$ and set $y^{(0)} = x^{(0)}$; repeat for $k = 1, 2, \dots$

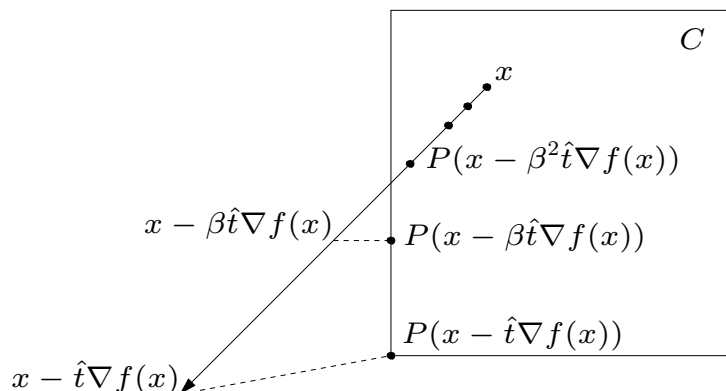
$$\begin{aligned} x^{(k)} &= P\left(y^{(k-1)} - t_k \nabla f(y^{(k-1)})\right) \\ y^{(k)} &= x^{(k)} + \frac{k-1}{k+2}(x^{(k)} - x^{(k-1)}) \end{aligned}$$

(note that $y^{(k)}$ is not necessarily feasible)

Backtracking line search

start at $t = \hat{t}$; repeat $t := \beta t$ until

$$f(x - tG_t(x)) \leq f(x) - t\nabla f(x)^T G_t(x) + \frac{t}{2} \|G_t(x)\|_2^2$$



- can take \hat{t} constant, or use $\hat{t} = t_{k-1}$ in the line search to determine t_k
- there exist other common variations of backtracking line search

Assumptions

- C closed and convex
- f convex with $\text{dom } f = \mathbf{R}^n$
- $\nabla f(x)$ is Lipschitz continuous with constant $L > 0$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \forall x, y$$

recall from lecture 1 that this implies

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|_2^2 \quad \forall x, y$$

Quadratic upper bound

for all x , and all $z \in C$,

$$f(x - tG_t(x)) \leq f(z) + G_t(x)^T(x - z) - t\left(1 - \frac{tL}{2}\right)\|G_t(x)\|_2^2$$

proof. define $v = G_t(x) - \nabla f(x)$

$$\begin{aligned} f(x - tG_t(x)) &\leq f(x) - t\nabla f(x)^T G_t(x) + \frac{Lt^2}{2}\|G_t(x)\|_2^2 \\ &= f(x) + tv^T G_t(x) - t\left(1 - \frac{tL}{2}\right)\|G_t(x)\|_2^2 \\ &\leq f(z) + \nabla f(x)^T(x - z) + tv^T G_t(x) - t\left(1 - \frac{tL}{2}\right)\|G_t(x)\|_2^2 \\ &\leq f(z) + G_t(x)^T(x - z) - t\left(1 - \frac{tL}{2}\right)\|G_t(x)\|_2^2 \end{aligned}$$

last line follows from $v^T(z - x + tG_t(x)) \leq 0$ (see page 5-9)

Analysis of projected gradient method

from quadratic upper bound, if $0 < t \leq 1/L$:

$$\begin{aligned} f(x - tG_t(x)) &\leq f^* + G_t(x)^T(x - x^*) - \frac{t}{2}\|G_t(x)\|_2^2 \\ &= f^* + \frac{1}{2t}(\|x - x^*\|_2^2 - \|x - tG_t(x) - x^*\|_2^2) \end{aligned}$$

as in lecture 1 (page 8), this implies

- with fixed step size t

$$f(x^{(k)}) - f^* \leq \frac{1}{2kt}\|x^{(0)} - x^*\|_2^2$$

- with backtracking line search, replace t with t_{\min}

Analysis of Nesterov's projected gradient method

from quadratic upper bound, if $0 < t \leq 1/L$ and $0 \leq \theta \leq 1$:

$$f(y - tG_t(y)) \leq f(x) + G_t(y)^T(y - x) - \frac{t}{2}\|G_t(y)\|_2^2$$

$$f(y - tG_t(y)) \leq f^* + G_t(y)^T(y - x^*) - \frac{t}{2}\|G_t(y)\|_2^2$$

$$f(y - tG_t(y)) \leq (1 - \theta)f(x) + \theta f^* + G_t(y)^T(y - (1 - \theta)x - \theta x^*) - \frac{t}{2}\|G_t(y)\|_2^2$$

rest of the proof is identical to lecture 1 (page 15)

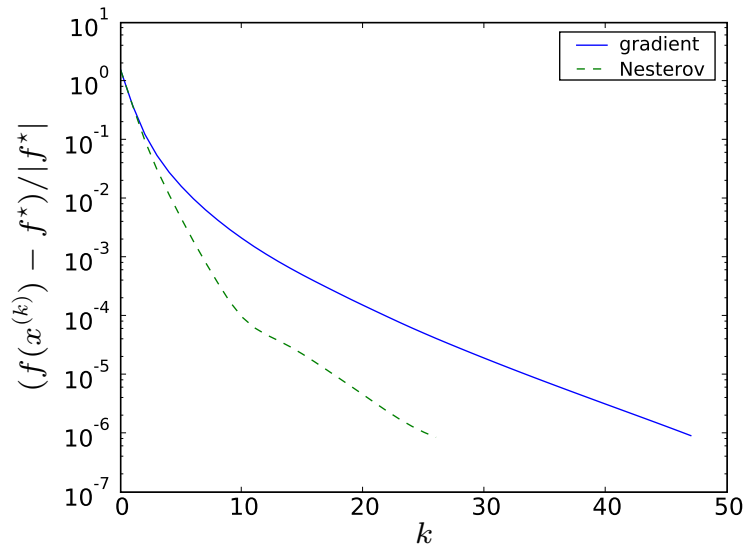
- with fixed step size t

$$f(x^{(k)}) - f^* \leq \frac{2}{(k + 1)^2 t}\|x^{(0)} - x^*\|_2^2$$

- with backtracking line search, replace t with t_{\min}

Example: QP with box constraints

$$\begin{aligned} &\text{minimize} && (1/2)x^T A x + b^T x \\ &\text{subject to} && 0 \preceq x \preceq \mathbf{1} \end{aligned}$$



$n = 3000$; fixed step size $t = 1/\lambda_{\max}(A)$

Gradient projection

5-17

Subgradient projection

$$x^{(k)} = P \left(x^{(k-1)} - t_k g^{(k-1)} \right)$$

with $g^{(k-1)} \in \partial f(x^{(k-1)})$

convergence results

- for constant step size, converges to neighborhood of optimal
- for diminishing nonsummable step sizes, converges

key idea in extending proof from lecture 3: projection on convex set does not increase distance between points

Gradient projection

5-18

Dual gradient methods

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \dots, m \end{array} \qquad \begin{array}{ll} \text{maximize} & g(\lambda) \\ \text{subject to} & \lambda \succeq 0 \end{array}$$

subgradient of negative dual function

$$-g(\lambda) = -\inf_x L(x, \lambda) = -\inf_x \left(f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) \right)$$

- if \hat{x} minimizes $L(\cdot, \lambda)$, then

$$-(f_1(\hat{x}), f_2(\hat{x}), \dots, f_m(\hat{x}))$$

is a subgradient of $-g$ at λ

- dual function is differentiable if \hat{x} is unique (e.g., f_0 is strictly convex)

Quadratic program

$$\begin{array}{ll} \text{minimize} & (1/2)x^T A x + b^T x \\ \text{subject to} & Cx \preceq d \end{array}$$

assume $A \succ 0$

dual problem

$$\begin{array}{ll} \text{maximize} & -(1/2)(C^T \lambda + b)^T A^{-1} (C^T \lambda + b) - d^T \lambda \\ \text{subject to} & \lambda \succeq 0 \end{array}$$

dual gradient method

$$x^{(k)} = -A^{-1} \left(C^T \lambda^{(k-1)} + b \right), \quad \lambda^{(k)} = \left(\lambda^{(k-1)} - t_k (d - Cx^{(k)}) \right)_+$$

primal iterates are not necessarily feasible

Dual decomposition

$$\begin{aligned} & \text{minimize} && \sum_{j=1}^p f_{0j}(x_j) \\ & \text{subject to} && \sum_{j=1}^p f_{ij}(x_j) \leq 0, \quad i = 1, \dots, m \end{aligned}$$

objective and constraint functions are separable functions of p variables x_j

dual function

$$g(\lambda) = \sum_{j=1}^p g_j(\lambda), \quad g_j(\lambda) \triangleq \inf_{x_j} \left(f_{0j}(x_j) + \sum_{i=1}^m \lambda_i f_{ij}(x_j) \right)$$

we will assume that the optimum in g_j is unique and attained for all $\lambda \succeq 0$

dual gradient method

- solve p unconstrained (and independent) subproblems

$$x_j^{(k)} = \operatorname{argmin}_{x_j} \left(f_{0j}(x_j) + \sum_{i=1}^m \lambda_i^{(k-1)} f_{ij}(x_j) \right), \quad j = 1, \dots, p$$

- make gradient update of λ

$$\lambda_i^{(k)} = \left(\lambda_i^{(k-1)} + t_k \sum_{j=1}^p f_{ij}(x_j^{(k)}) \right)_+$$

gives iterative decomposition of primal problem in p decoupled problems

interpretation: price coordination

- p units in a system; unit j chooses decision variable x_j
- constraints are limits on shared resources; λ_i is price of resource i
- the dual update
 - increases price λ_i if resource is over-utilized ($\sum_j f_{ij}(x_j) > 0$)
 - decreases price λ_i if resource is under-utilized ($\sum_j f_{ij}(x_j) < 0$)
 - never lets prices get negative

distributed architecture

- central node 0 sets price λ
- peripheral node j sets x_j

