ORIE 6326: Convex Optimization

Subgradient Method

Professor Udell

Operations Research and Information Engineering
Cornell

April 25, 2017

Some slides adapted from Stanford EE364b

Subgradient

recall: the subgradient generalizes the gradient for nonsmooth functions

for $f : \mathbf{R}^n \to \mathbf{R}$, we say $g \in \mathbf{R}^n$ is a subgradient of f at $x \in \mathbf{dom}(f)$ if

$$f(y) \geq f(x) + g^T(y - x) \quad \forall y \in \mathbf{dom}(f)$$

this is just the first-order condition for convexity

notation:

- $\partial f(x)$ denotes the set of all subgradients of f at x
- $\tilde{\nabla} f(x)$ denotes a particular choice of an element of $\partial f(x)$

Outline

Subgradients and optimality

Subgradients and descent

Subgradient method

Analysis of SGM

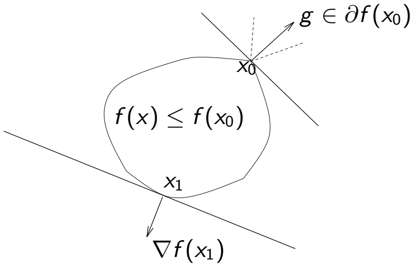
Examples

Polyak's optimal step size

Subgradients and sublevel sets

g is a subgradient at x means $f(y) \geq f(x) + g^T(y - x)$

hence $f(y) \leq f(x) \implies g^T(y - x) \leq 0$



- f differentiable at x_0: $\nabla f(x_0)$ is normal to the sublevel set $\{x \mid f(x) \leq f(x_0)\}$
- f nondifferentiable at x_0: subgradient defines a supporting hyperplane to sublevel set through x_0

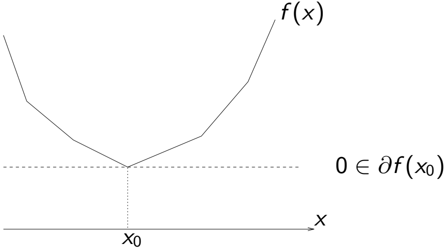
Optimality conditions — unconstrained

recall for f convex, differentiable,

$$f(x^\star) = \inf_x f(x) \iff 0 = \nabla f(x^\star)$$

generalization to nondifferentiable convex f:

$$f(x^\star) = \inf_x f(x) \iff 0 \in \partial f(x^\star)$$



proof.

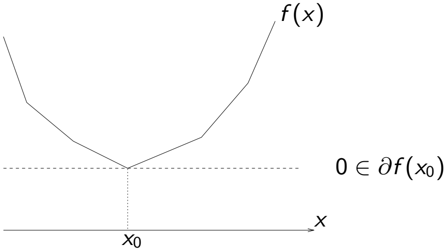
Optimality conditions — unconstrained

recall for f convex, differentiable,

$$f(x^\star) = \inf_x f(x) \iff 0 = \nabla f(x^\star)$$

generalization to nondifferentiable convex f:

$$f(x^\star) = \inf_x f(x) \iff 0 \in \partial f(x^\star)$$



proof. by definition (!)

$$f(y) \geq f(x^\star) + 0^T(y - x^\star) \text{ for all } y \iff 0 \in \partial f(x^\star)$$

...seems trivial but isn't

Example: piecewise linear minimization

$f(x) = \max_{i=1,\ldots,m}(a_i^T x + b_i)$

$x^\star$ minimizes f $\iff 0 \in \partial f(x^\star) = \mathbf{conv}\{a_i \mid a_i^T x^\star + b_i = f(x^\star)\}$

$\iff$ there is a λ with

$$\lambda \succeq 0, \qquad \mathbf{1}^T\lambda = 1, \qquad \sum_{i=1}^m \lambda_i a_i = 0$$

where $\lambda_i = 0$ if $a_i^T x^\star + b_i < f(x^\star)$

...but these are the KKT conditions for the epigraph form

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & a_i^T x + b_i \leq t, \quad i = 1, \ldots, m \end{array}$$

with dual

$$\begin{array}{ll} \text{maximize} & b^T \lambda \\ \text{subject to} & \lambda \succeq 0, \qquad A^T \lambda = 0, \qquad \mathbf{1}^T \lambda = 1 \end{array}$$

Optimality conditions — constrained

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

we assume

- f_i convex, defined on $\mathbf{R}^n$ (hence subdifferentiable)
- strict feasibility (Slater's condition)

$x^\star$ is primal optimal ($\lambda^\star$ is dual optimal) iff

$$\begin{aligned} & f_i(x^\star) \le 0, \quad \lambda_i^\star \ge 0 \\ & 0 \in \partial f_0(x^\star) + \textstyle\sum_{i=1}^m \lambda_i^\star \partial f_i(x^\star) \\ & \lambda_i^\star f_i(x^\star) = 0 \end{aligned}$$

...generalizes KKT for nondifferentiable f_i

Outline

Subgradients and optimality

Subgradients and descent

Subgradient method

Analysis of SGM

Examples

Polyak's optimal step size

Directional derivative

directional derivative of f at x in the direction δx is

$$f'(x;\delta x) \triangleq \lim_{h \searrow 0} \frac{f(x+h\delta x)-f(x)}{h}$$

can be $+\infty$ or $-\infty$

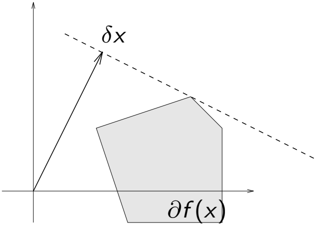
- f convex, finite near $x \implies f'(x;\delta x)$ exists
- f differentiable at x iff, for some g $(=\nabla f(x))$ and all δx,

$$f'(x;\delta x) = g^T \delta x$$

  (*i.e.*, $f'(x;\delta x)$ is a linear function of δx)

Directional derivative and subdifferential

general formula for convex f: $f'(x;\delta x) = \sup_{g \in \partial f(x)} g^T \delta x$



Descent directions

δx is a **descent direction** for f at x if $f'(x; \delta x) < 0$

for differentiable f, $\delta x = -\nabla f(x)$ is always a descent direction (except when it is zero)

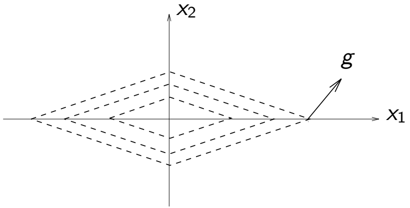
Descent directions

δx is a **descent direction** for f at x if $f'(x; \delta x) < 0$

for differentiable f, $\delta x = -\nabla f(x)$ is always a descent direction (except when it is zero)

warning: for nondifferentiable (convex) functions, $\tilde{\nabla} f(x)$ need not be descent direction

example: $f(x) = |x_1| + 2|x_2|$



Subgradients and distance to sublevel sets

if f is convex, $f(z) < f(x)$, $g \in \partial f(x)$, then for small enough $t > 0$,

$$\|x - tg - z\|_2 < \|x - z\|_2$$

thus $-g$ is descent direction for $\|x - z\|_2$, for **any** z with $f(z) < f(x)$ (*e.g.*, $x^\star$)

negative subgradient is descent direction for distance to optimal point

proof:

Subgradients and distance to sublevel sets

if f is convex, $f(z) < f(x)$, $g \in \partial f(x)$, then for small enough $t > 0$,

$$\|x - tg - z\|_2 < \|x - z\|_2$$

thus $-g$ is descent direction for $\|x - z\|_2$, for **any** z with $f(z) < f(x)$ (*e.g.*, $x^\star$)

negative subgradient is descent direction for distance to optimal point

proof:
$$\begin{aligned} \|x - tg - z\|_2^2 &= \|x - z\|_2^2 - 2tg^T(x - z) + t^2\|g\|_2^2 \\ &\leq \|x - z\|_2^2 - 2t(f(x) - f(z)) + t^2\|g\|_2^2 \end{aligned}$$

Descent directions and optimality

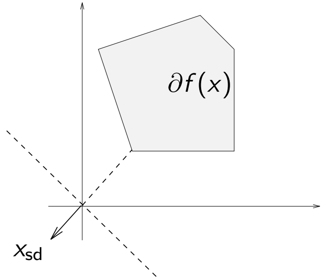
fact: for f convex, finite near x, either

- $0 \in \partial f(x)$ (in which case x minimizes f), or
- there is a descent direction for f at x

i.e., x is optimal (minimizes f) iff there's no descent direction for f at x

proof: define $\delta x_{\mathrm{sd}} = -\underset{z \in \partial f(x)}{\operatorname{argmin}} \|z\|_2$

if $\delta x_{\mathrm{sd}} = 0$, then $0 \in \partial f(x)$, so x is optimal; otherwise $f'(x; \delta x_{\mathrm{sd}}) = -\left(\inf_{z \in \partial f(x)} \|z\|_2\right)^2 < 0$, so δx_{sd} is a descent direction



Outline

Subgradient method

the **subgradient method** is a simple algorithm to minimize nondifferentiable convex function f

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$$

- $x^{(k)}$ is the kth iterate
- $g^{(k)}$ is **any** subgradient of f at $x^{(k)}$
- $\alpha_k > 0$ is the kth step size

Subgradient method

the **subgradient method** is a simple algorithm to minimize nondifferentiable convex function f

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$$

- $x^{(k)}$ is the kth iterate
- $g^{(k)}$ is **any** subgradient of f at $x^{(k)}$
- $\alpha_k > 0$ is the kth step size

warning: subgradient method is **not** a descent method.
so, do **not** call it subgradient descent [Steven Wright].

Subgradient method

the **subgradient method** is a simple algorithm to minimize nondifferentiable convex function f

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$$

- $x^{(k)}$ is the kth iterate
- $g^{(k)}$ is **any** subgradient of f at $x^{(k)}$
- $\alpha_k > 0$ is the kth step size

warning: subgradient method is **not** a descent method.
so, do **not** call it subgradient descent [Steven Wright].

instead, keep track of best point so far

$$f_{\text{best}}^{(k)} = \min_{i=1,\ldots,k} f(x^{(i)})$$

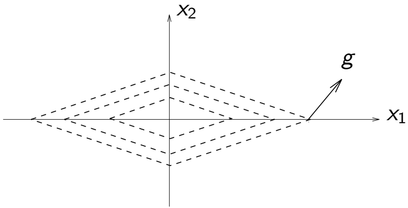
Line search + subgradients = danger

warning: with exact line search, subgradient method can converge to suboptimal point!

Line search + subgradients = danger

warning: with exact line search, subgradient method can converge to suboptimal point!

example: $f(x) = |x_1| + 2|x_2|$



Step size rules

step sizes are fixed ahead of time (can't do linesearch)

- **constant step size:** $\alpha_k = \alpha$ (constant)
- **constant step length:** $\alpha_k = \gamma / \|g^{(k)}\|_2$ (so $\|x^{(k+1)} - x^{(k)}\|_2 = \gamma$)
- **square summable but not summable:** step sizes satisfy

$$\sum_{k=1}^{\infty} \alpha_k^2 < \infty, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty$$

  e.g., $\alpha_k = \frac{1}{k}$

- **nonsummable diminishing:** step sizes satisfy

$$\lim_{k \to \infty} \alpha_k = 0, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty,$$

  e.g., $\alpha_k = \frac{1}{\sqrt{k}}$

Outline

Subgradients and optimality

Subgradients and descent

Subgradient method

Analysis of SGM

Examples

Polyak's optimal step size

Lipschitz functions

a function $f : \mathbf{R}^n \to \mathbf{R}$ is L-**Lipschitz continuous** if for every $x, y \in \mathbf{dom}\, f$,

$$|f(x) - f(y)| \le L\|x - y\|$$

(often, we simply say f is L-Lipschitz)

Lipschitz functions

a function $f : \mathbf{R}^n \to \mathbf{R}$ is L-**Lipschitz continuous** if for every $x, y \in \textbf{dom} f$,

$$|f(x) - f(y)| \leq L\|x - y\|$$

(often, we simply say f is L-Lipschitz)

notice for any $g \in \partial f(x)$,

$$f(y) - f(x) \leq g^T(y - x) \leq \|g\|\|y - x\|$$

so the Lipschitz constant for f satisfies

$$L \leq \sup_{x \in \textbf{dom} f} \sup_{g \in \partial f(x)} \|g\|$$

Lipschitz functions

examples:

- $|x|$:

Lipschitz functions

examples:

- $|x|$: $L = 1$

Lipschitz functions

examples:

- $|x|$: $L = 1$
- $\max(x, 0)$:

Lipschitz functions

examples:

- $|x|$: $L = 1$
- $\max(x, 0)$: $L = 1$

Lipschitz functions

examples:

- $|x|$: $L = 1$
- $\max(x, 0)$: $L = 1$
- $\|x\|_1$:

Lipschitz functions

examples:

- $|x|$: $L = 1$
- $\max(x, 0)$: $L = 1$
- $\|x\|_1$: $L = \sqrt(n)$ (subgradients are sign vectors)

Lipschitz functions

examples:

- $|x|$: $L = 1$
- $\max(x, 0)$: $L = 1$
- $\|x\|_1$: $L = \sqrt(n)$ (subgradients are sign vectors)
- $\|x\|^2$ for $\|x\| \leq R$:

Lipschitz functions

examples:

- $|x|$: $L = 1$
- $\max(x, 0)$: $L = 1$
- $\|x\|_1$: $L = \sqrt(n)$ (subgradients are sign vectors)
- $\|x\|^2$ for $\|x\| \leq R$: $L = R$

Lipschitz functions

examples:

- $|x|$: $L = 1$
- $\max(x, 0)$: $L = 1$
- $\|x\|_1$: $L = \sqrt(n)$ (subgradients are sign vectors)
- $\|x\|^2$ for $\|x\| \leq R$: $L = R$
- $\lambda_{\max}(X)$:

Lipschitz functions

examples:

- $|x|$: $L = 1$
- $\max(x, 0)$: $L = 1$
- $\|x\|_1$: $L = \sqrt(n)$ (subgradients are sign vectors)
- $\|x\|^2$ for $\|x\| \leq R$: $L = R$
- $\lambda_{\max}(X)$: $L = 1$ (subgradients are eigenvectors)

Lipschitz functions

examples:

- $|x|$: $L = 1$
- $\max(x, 0)$: $L = 1$
- $\|x\|_1$: $L = \sqrt(n)$ (subgradients are sign vectors)
- $\|x\|^2$ for $\|x\| \leq R$: $L = R$
- $\lambda_{\max}(X)$: $L = 1$ (subgradients are eigenvectors)
- indicators of convex sets $\mathcal{C}$:

Lipschitz functions

examples:

- $|x|$: $L = 1$
- $\max(x, 0)$: $L = 1$
- $\|x\|_1$: $L = \sqrt(n)$ (subgradients are sign vectors)
- $\|x\|^2$ for $\|x\| \leq R$: $L = R$
- $\lambda_{\max}(X)$: $L = 1$ (subgradients are eigenvectors)
- indicators of convex sets $\mathcal{C}$: not Lipschitz continuous

Lipschitz functions

examples:

- $|x|$: $L = 1$
- $\max(x, 0)$: $L = 1$
- $\|x\|_1$: $L = \sqrt(n)$ (subgradients are sign vectors)
- $\|x\|^2$ for $\|x\| \leq R$: $L = R$
- $\lambda_{\max}(X)$: $L = 1$ (subgradients are eigenvectors)
- indicators of convex sets $\mathcal{C}$: not Lipschitz continuous

Assumptions

- f is bounded below: $p^\star = \inf_x f(x) > -\infty$
- f attains its minimum at $x^\star$: $f(x^\star) = p^\star$
- f is L-Lipschitz, so

$$\|g\|_2 \leq L \quad \text{for all} \quad g \in \partial f$$

- $\|x^{(0)} - x^\star\|_2 \leq R$

these assumptions are stronger than needed, just to simplify proofs

Convergence results

define

- averaged iterate $\bar{x}^{(k)} = \frac{1}{k}\sum_{i=1}^{k} x^{(k)}$
- averaged value $\bar{f}^{(k)} = f(\bar{x}^{(k)})$
- $\bar{f} = \lim_{k\to\infty} \bar{f}^{(k)}$
- $f_{\rm best} = \lim_{k\to\infty} \bar{f}^{(k)}_{\rm best}$

we'll show $\bar{f}^{(k)} \to \bar{f}$ is close to $p^\star$.

- notice $f^{(k)}_{\rm best} \le \bar{f}^{(k)}$
- advantage of averaging: no need to evaluate f!

- **constant step size:** $\bar{f} - p^\star \le L^2\alpha/2$, *i.e.*, **converges to** $L^2\alpha/2$**-suboptimal** (converges to $p^\star$ if f differentiable, α small enough)

- **constant step length:** $f_{\rm best} - p^\star \le L\gamma/2$, *i.e.*, **converges to** $L\gamma/2$**-suboptimal**

- **diminishing step size rule:** $f_{\rm best} = p^\star$, *i.e.*, **converges**

Subgradient method decreases distance to optimal set

key quantity: Euclidean distance to the optimal set,
not the function value

let $x^\star$ be any minimizer of f

$$
\begin{aligned}
\|x^{(k+1)} - x^\star\|_2^2 &= \|x^{(k)} - \alpha_k g^{(k)} - x^\star\|_2^2 \\
&= \|x^{(k)} - x^\star\|_2^2 - 2\alpha_k g^{(k)T}(x^{(k)} - x^\star) + \alpha_k^2 \|g^{(k)}\|_2^2 \\
&\leq \|x^{(k)} - x^\star\|_2^2 - 2\alpha_k (f(x^{(k)}) - p^\star) + \alpha_k^2 \|g^{(k)}\|_2^2
\end{aligned}
$$

using $p^\star = f(x^\star) \geq f(x^{(k)}) + g^{(k)T}(x^\star - x^{(k)})$

apply recursively to get

$$
\begin{aligned}
\|x^{(k+1)} - x^\star\|_2^2 &\leq \|x^{(1)} - x^\star\|_2^2 - 2\sum_{i=1}^{k} \alpha_i (f(x^{(i)}) - p^\star) + \sum_{i=1}^{k} \alpha_i^2 \|g^{(i)}\|_2^2 \\
&\leq R^2 - 2\sum_{i=1}^{k} \alpha_i (f(x^{(i)}) - p^\star) + L^2 \sum_{i=1}^{k} \alpha_i^2
\end{aligned}
$$

Subgradient method for constant step size

for constant step size $\alpha_k = \alpha$, we can use

$$\begin{aligned}\sum_{i=1}^{k} \alpha(f(x^{(i)}) - p^\star) &= (\alpha \sum_{i=1}^{k} f(x^{(i)})) - \alpha k p^\star \\ &\geq \alpha k f(\bar{x}^{(k)} - p^\star) = \alpha k f(\bar{x}^{(k)} - p^\star)\end{aligned}$$

(where the inequality is Jensen's) to get

$$\bar{f}^{(k)} - p^\star \leq \frac{R^2}{2k\alpha} + L^2\alpha/2.$$

- righthand side converges to $L^2\alpha/2$ as $k \to \infty$
- for fixed k, choose $\alpha = \frac{R}{L\sqrt{k}}$ to minimize bound

$$\bar{f}^{(k)} - p^\star \leq \frac{RL}{2\sqrt{k}} + \frac{RL}{2\sqrt{k}} = \frac{RL}{\sqrt{k}}.$$

Subgradient method for varying step size

by recursive application of subgradient inequality, we had

$$\begin{aligned}\|x^{(k+1)} - x^\star\|_2^2 &\leq \|x^{(1)} - x^\star\|_2^2 - 2\sum_{i=1}^{k} \alpha_i (f(x^{(i)}) - p^\star) + \sum_{i=1}^{k} \alpha_i^2 \|g^{(i)}\|_2^2 \\ &\leq R^2 - 2\sum_{i=1}^{k} \alpha_i (f(x^{(i)}) - p^\star) + L^2 \sum_{i=1}^{k} \alpha_i^2\end{aligned}$$

so for changing step size, use

$$\sum_{i=1}^{k} \alpha_i (f(x^{(i)}) - p^\star) \geq (f_{\mathrm{best}}^{(k)} - p^\star) \left(\sum_{i=1}^{k} \alpha_i\right)$$

to get

$$f_{\mathrm{best}}^{(k)} - p^\star \leq \frac{R^2 + L^2 \sum_{i=1}^{k} \alpha_i^2}{2\sum_{i=1}^{k} \alpha_i}.$$

Subgradient method for varying step size

constant step length: for $\alpha_k = \gamma/\|g^{(k)}\|_2$ we get

$$f_{\mathrm{best}}^{(k)} - p^\star \leq \frac{R^2 + \sum_{i=1}^k \alpha_i^2 \|g^{(i)}\|_2^2}{2\sum_{i=1}^k \alpha_i} \leq \frac{R^2 + \gamma^2 k}{2\gamma k/L},$$

righthand side converges to $L\gamma/2$ as $k \to \infty$

square summable but not summable step sizes:
suppose step sizes satisfy

$$\sum_{i=1}^{\infty} \alpha_k^2 < \infty, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty$$

then

$$f_{\mathrm{best}}^{(k)} - p^\star \leq \frac{R^2 + L^2 \sum_{i=1}^k \alpha_i^2}{2\sum_{i=1}^k \alpha_i}$$

as $k \to \infty$, numerator converges to a finite number, denominator converges to ∞, so $f_{\mathrm{best}}^{(k)} \to p^\star$

Stopping criterion

- terminating when $\dfrac{R^2 + L^2 \sum_{i=1}^k \alpha_i^2}{2 \sum_{i=1}^k \alpha_i} \leq \epsilon$ is really, really, slow
- we saw that if $\alpha_i = \alpha = \frac{R}{L\sqrt{k}}$ then after k steps, $\epsilon \leq \frac{RL}{\sqrt{k}}$, so for accuracy ϵ need

  $$k \geq \frac{R^2 L^2}{\epsilon^2}$$

  iterations
- the truth: there's no good stopping criterion for the subgradient method . . .

Outline

Example: Piecewise linear minimization

$$\text{minimize} \quad f(x) = \max_{i=1,\ldots,m}(a_i^T x + b_i)$$

to find a subgradient of f: find index j for which

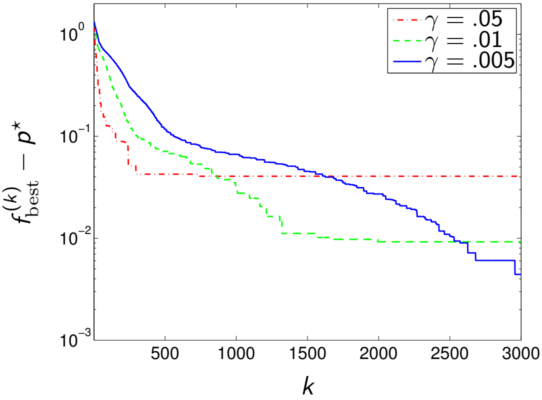
$$a_j^T x + b_j = \max_{i=1,\ldots,m}(a_i^T x + b_i)$$

and take $g = a_j$

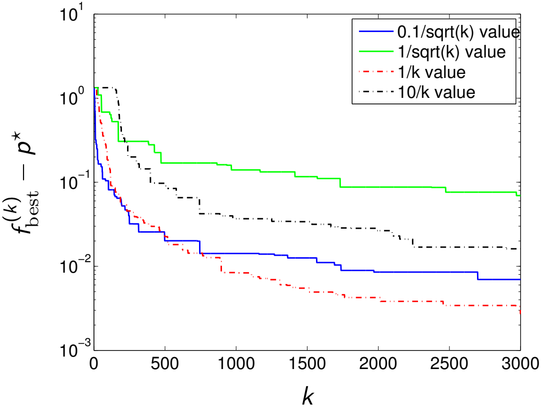
subgradient method: $x^{(k+1)} = x^{(k)} - \alpha_k a_j$

problem instance with $n = 20$ variables, $m = 100$ terms, $p^\star \approx 1.1$

$f_{\text{best}}^{(k)} - p^\star$, constant step length $\gamma = 0.05, 0.01, 0.005$



diminishing step rules $\alpha_k = 0.1/\sqrt{k}$ and $\alpha_k = 1/\sqrt{k}$, square summable step size rules $\alpha_k = 1/k$ and $\alpha_k = 10/k$



How to avoid slow convergence

don't use subgradient method when you want very high accuracy!

instead,

- for highest accuracy: rewrite problem as LP or SDP; use IPM
- for medium accuracy:
  - regularize your objective (so it's strongly convex)

    $$\tilde{f}(x) = f(x) + \alpha \|x - x^0\|^2$$

  - smooth your objective (so it's smooth)

    $$\tilde{f}(x) = \mathop{\mathbf{E}}_{y:\|y-x\|\leq\delta} f(y)$$

  - more on these later...
- for low accuracy: use a constant step size; terminate when you stop improving much or get bored

Outline

Optimal step size when $p^\star$ is known

- choice due to Polyak:

$$\alpha_k = \frac{f(x^{(k)}) - p^\star}{\|g^{(k)}\|_2^2}$$

(can also use when optimal value is estimated)

- motivation: start with basic inequality

$$\|x^{(k+1)} - x^\star\|_2^2 \leq \|x^{(k)} - x^\star\|_2^2 - 2\alpha_k(f(x^{(k)}) - p^\star) + \alpha_k^2\|g^{(k)}\|_2^2$$

and choose α_k to minimize righthand side

- yields

$$\|x^{(k+1)} - x^\star\|_2^2 \le \|x^{(k)} - x^\star\|_2^2 - \frac{(f(x^{(k)}) - p^\star)^2}{\|g^{(k)}\|_2^2}$$

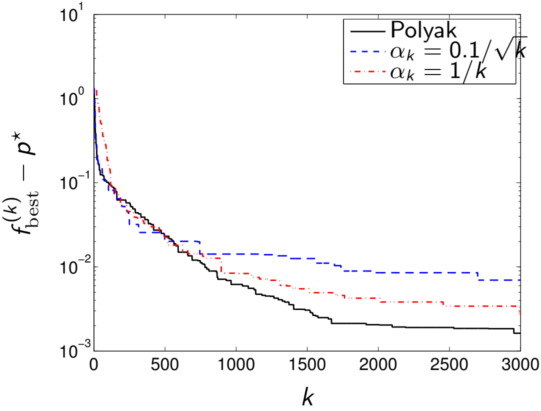
(in particular, $\|x^{(k)} - x^\star\|_2$ decreases each step)
- applying recursively,

$$\sum_{i=1}^{k} \frac{(f(x^{(i)}) - p^\star)^2}{\|g^{(i)}\|_2^2} \le R^2$$

and so using Jensen twice,

$$\begin{aligned} k\left(\frac{1}{k}\sum_{i=1}^{k} f(x^{(i)}) - p^\star\right)^2 \le \sum_{i=1}^{k}(f(x^{(i)}) - p^\star)^2 &\le R^2L^2 \\ \frac{1}{k}\sum_{i=1}^{k} f(x^{(i)}) - p^\star &\le \frac{RL}{\sqrt{k}} \\ f(\bar{x}^{(i)}) - p^\star &\le \frac{RL}{\sqrt{k}} \end{aligned}$$

PWL example with Polyak's step size, $\alpha_k = 0.1/\sqrt{k}$, $\alpha_k = 1/k$



Finding a point in the intersection of convex sets

$C = C_1 \cap \cdots C_m$ is nonempty, $C_1, \ldots, C_m \subseteq \mathbf{R}^n$ closed and convex

find a point in C by minimizing

$$f(x) = \max\{\mathbf{dist}(x, C_1), \ldots, \mathbf{dist}(x, C_m)\}$$

with $\mathbf{dist}(x, C_j) = f(x)$, a subgradient of f is

$$g = \nabla\, \mathbf{dist}(x, C_j) = \frac{x - P_{C_j}(x)}{\|x - P_{C_j}(x)\|_2}$$

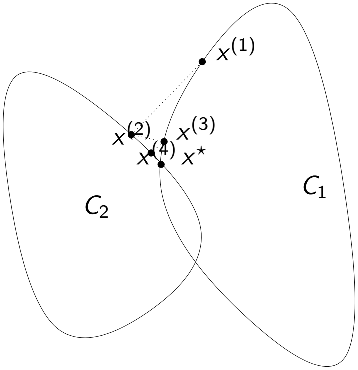
subgradient update with optimal step size:

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - \alpha_k g^{(k)} \\ &= x^{(k)} - f(x^{(k)}) \frac{x - P_{C_j}(x)}{\|x - P_{C_j}(x)\|_2} \\ &= P_{C_j}(x^{(k)}) \end{aligned}$$

- a version of the famous **alternating projections** algorithm
- at each step, project the current point onto the farthest set
- for $m = 2$ sets, projections alternate onto one set, then the other
- convergence: $\mathbf{dist}(x^{(k)}, C) \to 0$ as $k \to \infty$

Alternating projections

first few iterations:



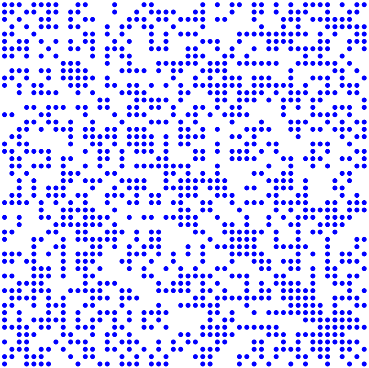
$\ldots x^{(k)}$ eventually converges to a point $x^\star \in C_1 \cap C_2$

Example: Positive semidefinite matrix completion

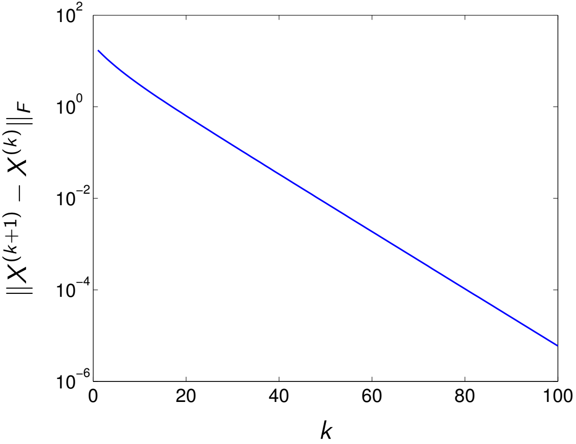
- some entries of matrix in $\mathbf{S}^n$ fixed
- find values for others so completed matrix is PSD
- $C_1 = \mathbf{S}^n_+$, C_2 is (affine) set in $\mathbf{S}^n$ with specified fixed entries
- projection onto C_1 by eigenvalue decomposition, truncation: for $X = \sum_{i=1}^n \lambda_i q_i q_i^T$,

$$P_{C_1}(X) = \sum_{i=1}^n \max\{0, \lambda_i\} q_i q_i^T$$

- projection of X onto C_2 by setting specified entries to fixed values specific example: 50×50 matrix missing about half of its entries



- initialize $X^{(1)}$ with unknown entries set to 0 convergence is linear:



Polyak step size when $p^\star$ isn't known

- use step size
$$\alpha_k = \frac{f(x^{(k)}) - f_{\text{best}}^{(k)} + \gamma_k}{\|g^{(k)}\|_2^2}$$
with $\sum_{k=1}^{\infty} \gamma_k = \infty$, $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$
- $f_{\text{best}}^{(k)} - \gamma_k$ serves as estimate of $p^\star$
- γ_k is in scale of objective value
- can show $f_{\text{best}}^{(k)} \to p^\star$

PWL example with Polyak's step size, using $p^\star$, and estimated with $\gamma_k = 10/(10+k)$