- Project subgradient method
  solves constrained optimization problem

$$\begin{aligned} &\min. \ f(x) && f: \mathbb{R}^n \to \mathbb{R} \text{ is convex} \\ &\text{s.t.} \ \ x \in C && C \subseteq \mathbb{R}^n \text{ is convex} \end{aligned}$$

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

projection does not increase the distance to x^*

(once you have a merit function, you can modify your algorithm way as long as whatever you do does not increase that function

e.g. linear equality constraints

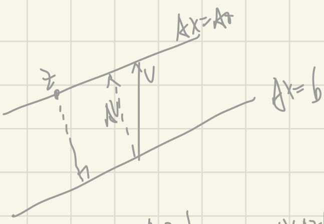
$$\begin{aligned} &\min. \ f(x) \\ &\text{s.t.} \ \ Ax = b \end{aligned}$$

projection z onto $\{x \mid Ax = b\}$

$$\begin{aligned} p(z) &= z - A^T (AA^T)^{-1} (Az - b) \\ &= [I - A^T (AA^T)^{-1} A] z + A (AA^T)^{-1} b \end{aligned}$$

projected subgradient update is

$$\begin{aligned} x^{(k+1)} &= P(x^{(k)} - a_k g^{(k)}) \\ &= x^{(k)} - \alpha_k P(g^{(k)}) \\ &= x^{(k)} - \alpha_k [I - A^T (AA^T)^{-1} A] g^{(k)} \\ &= x^{(k)} - \alpha_k P_{N(A)}(g^{k}) \end{aligned}$$



$Av = Az - b$

$A^T (AA^T)^{-1} Av = A^T (AA^T)^{-1} (Az - b)$

$v = A^T (AA^T)^{-1} (Az - b)$

e.g. $\min \|x\|_1$
s.t. $Ax = b$

$$g(x) = \text{sign}(x)$$

$$x^{(k+1)} = x^{(k)} - a_k (I - A^T (AA^T)^{-1} A) \ \text{sign}(x)$$

· Projected subgradient for dual problem

convex primal problem.

$$\min.\ f(x)$$
$$s.t.\ f_i(x) \le 0$$

solve dual problem

$$\max\ g(\lambda)$$
$$s.t\ \lambda_i \ge 0$$

via projected subgradient method

$$\lambda^{(k+1)} = P(\lambda^{(k)} - \alpha_k h) \qquad h \in \partial(-g)(\lambda^{(k)})$$

when λ^* is found $\mathcal{L}(x, \lambda^*)$ should have a unique minimizer in x : x^*

one condition for this is that $f_0(x)$ is strictly convex ($f_0 + \sum_i \lambda_i f_i(x)$) is strictly convex)

· Subgradient of negative dual function

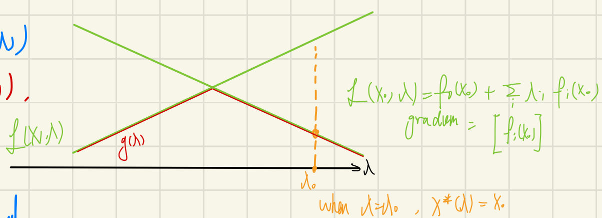
assume f_0 is strictly convex, $\lambda \ge 0$

$$x^*(\lambda) = \arg\min_x [f_0(x) + \sum_i \lambda_i f_i(x)]$$

so $g(\lambda) = \inf_x \mathcal{L}(x, \lambda) = \mathcal{L}(x^*(\lambda), \lambda) = f_0(x^*(\lambda)) + \sum_i \lambda_i \cdot f_i(x^*(\lambda))$

$$-g(\lambda) = -f_0(x^*(\lambda)) + \sum_i -\lambda_i \cdot f(x^*(\lambda))$$

a subgradient for $-g$ is $h_i = -f_i(x^*(\lambda))$.



projected subgradient method for the dual

$x^{(k)} = x^*(\lambda^{(k)})$ minimize the Lagrangian at current λ
(without considering the feasibility of the primal problem)

$$\mathcal{L}(x, \lambda) = f(x) + \underbrace{\sum_i \lambda_i f_i(x)}_{\text{cost of constraints}}$$

$\lambda_i^{(k+1)} = P(\lambda_i^{(k)} + \alpha_k f_i(x^{(k)}))$ // price update

if $f_i(x^{(k)}) > 0$, then $x^{(k)}$ is violating constraint i

means: when min the Lagrangian, the price for violating ith constraint is not high enough

if $f_i(x^{(k)}) < 0$, then its possible that in the final solution, ith constraint is slack

drop the price of ith resource since slackness not rewarded but never goes below 0 produce a lower bound $x^{(k)} = x^*(\lambda^{(k)})$ $g(\lambda^{(k)}) = L(x^{(k)}, \lambda^{(k)})$

produce a upper bound $f_0(\tilde{x}^{(k)})$ $\tilde{x}^{(k)}$ is projection of $x^{(k)}$ on feasible set

eg. minimizing quadratic function over the unit box

$$\begin{aligned} \min.\ & \tfrac{1}{2} x^T P x - q^T x \qquad P \succ 0 \\ \text{s.t.}\ & x_i^2 \le 1 \end{aligned}$$

$$L(x,\lambda) = \tfrac{1}{2} x^T P x - q^T x + x^T \operatorname{diag}(\lambda) x - \mathbf{1}^T \lambda$$

$$\nabla_x L(x,\lambda) = Px - q + 2\operatorname{diag}(\lambda) x := 0$$

$$x^*(\lambda) = [P + 2\operatorname{diag}(\lambda)]^{-1} q$$

$$\begin{cases} x^*(\lambda) = [P + 2\operatorname{diag}(\lambda)]^{-1} q \\ \lambda_i^{(k+1)} = [\lambda_i^{(k)} + \alpha_k \cdot ((x_i^{(k)})^2 - 1)]_+ \end{cases}$$

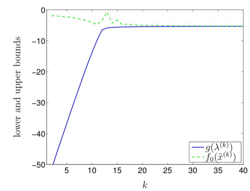


Figure 9: The values of the lower bound $g(\lambda^{(k)})$ and the upper bound $f_0(\tilde{x}^{(k)})$, versus the iteration number k. We use the fixed step size with $\alpha = 0.1$.

fixed step size for differentiable function (g)

· Subgradient method for constrained optimization

$$\begin{aligned} \min.\ & f_0(x) \qquad f_i : \mathbb{R}^n \to \mathbb{R} \text{ is convex} \\ \text{s.t.}\ & f_i(x) \le 0 \end{aligned}$$

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

$$g^{(k)} \in \begin{cases} \partial f_0(x) & f_i(x) \le 0 \\ \partial f_j(x) & f_j(x) > 0 \end{cases}$$

(f_j may even go larger since subgradient direction is not necessarily a descent direction)

$$f_{best}^{(k)} = \min\{ f_0(x^{(i)}) \mid x^{(i)} \text{ feasible},\ i = 1, \dots, m \}$$

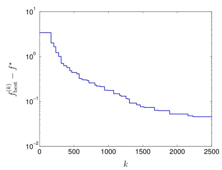


Figure 10: The value of $f_{\text{best}}^{(k)} - f^*$ versus the iteration number k. In this case, we use the square summable step size with $\alpha_k = 1/k$ for the optimality update.

can do Polyak step size, since we know $f_j(x) = 0$

- Noisy unbiased subgradient

random vector $\tilde{g} \in \mathbb{R}^n$ is a noisy unbiased subgradient for $f: \mathbb{R}^n \to \mathbb{R}$ at x if

$$f(z) \geq f(x) + E[\tilde{g}]^T (z-x)$$

for all z, i.e. $g = E[\tilde{g}] \in \partial f(x)$

same as: $\tilde{g} = g + v$ $E[v] = 0$ v can represent error in computing g, measurement noise, Monte Carlo sampling error, etc

if x is also random, $\tilde{g}$ is a noisy unbiased subgradient of f at x if

$$\forall z,\ f(z) \geq f(x) + E[\tilde{g}|x]^T (z-x)$$

holds almost surely

same as $E[\tilde{g}|x] \in \partial f(x)$ (almost surely)

- Stochastic subgradient method

Same as subgradient method, but using noisy unbiased subgradient

$$x^{(k+1)} = x^{(k)} - \alpha_k \cdot \tilde{g}^{(k)} \qquad E[\tilde{g}^{(k)} | x^{(k)}] = g^{(k)} \in \partial f(x^{(k)})$$

$g^{(k)}$ is any noisy unbiased subgradient of f at $x^{(k)}$

- Assumptions

$f^* = \inf_x f(x) > -\infty$ with $f(x^*) = f^*$

$E[\|g^{(k)}\|_2^2] \leq G^2$ for all k

$E[\|x^{(1)} - x^*\|_2^2] \leq R^2$

Step size are square-summable but not summable

$$a_k \geq 0 \quad \sum_k a_k^2 = \|a\|_2^2 < \infty \quad \sum_{k=1}^{\infty} a_k = \infty$$

- Convergence results

convergence in expectation: $\lim_{k \to \infty} E f_{best}^{(k)} = f^*$

convergence in probability: for any $\epsilon > 0$ $\lim_{k \to \infty} \text{Prob}(f_{best}^{(k)} \geq f^* + \epsilon) = 0$

almost sure convergence: $\lim_{k \to \infty} f_{best}^{(k)} = f^*$

· Convergence proof

key quantity: expected Euclidean distance to the optimal set

$$E\left(\|x^{(k+1)}-x^*\|_2^2 \mid x^{(k)}\right) = E\left(\|x^{(k)}-a_k\cdot\tilde{g}^{(k)}-x^*\|_2^2 \mid x^{(k)}\right)$$

$$= \|x^{(k)}-x^*\|_2^2 - 2a_k\cdot E\left[\tilde{g}^{(k)T}(x^{(k)}-x^*) \mid x^{(k)}\right] + a_k^2\cdot E\left[\|g^{(k)}\|_2^2 \mid x^{(k)}\right]$$

$$= \|x^{(k)}-x^*\|_2^2 - 2a_k\cdot(x^{(k)}-x^*)^T E(\tilde{g}^{(k)} \mid x^{(k)}) + a_k^2\cdot E\left[\|g^{(k)}\|_2^2 \mid x^{(k)}\right]$$

$$f(x^*) \le f(x^{(k)}) + E[\tilde{g}(x^{(k)})]^T(x^*-x^{(k)})$$

$$\le \|x^{(k)}-x^*\|_2^2 - 2a_k\left(f(x^{(k)})-f^*\right) + a_k^2\cdot E\left[\|g^{(k)}\|_2^2 \mid x^{(k)}\right]$$

$$\underset{x^{(k)}}{E}\left[\|x^{(k+1)}-x^*\|_2^2\right] \le \underset{x^{(k)}}{E}\left[\|x^{(k)}-x^*\|_2^2\right] - 2a_k\left(\underset{x^{(k)}}{E}[f(x^{(k)})]-f^*\right) + a_k^2\,\underset{x^{(k)}}{E}\left[\|g^{(k)}\|_2^2\right]$$

$$\vdots$$

$$\le \underset{x^{(k)}}{E}\left[\|x^{(1)}-x^*\|_2^2\right] - 2\sum_k a_k\left(\underset{x^{(k)}}{E}[f(x^{(k)})]-f^*\right) + \sum_k a_k^2\,\underset{x^{(k)}}{E}\left[\|g^{(k)}\|_2^2\right]$$

$$\le E\left[\|x^{(1)}-x^*\|_2^2\right] - 2\sum_k a_k\left(E[f(x^{(k)})]-f^*\right) + G^2\sum_k a_k^2$$

$$\le E\left[\|x^{(1)}-x^*\|_2^2\right] - 2\sum_k a_k\cdot\min_i\left\{E[f(x^{(i)})]-f^*\right\} + G^2\cdot\sum_k a_k^2$$

$$\min_i\left\{E[f(x^{(i)})]-f^*\right\} \le \frac{R^2+G^2\cdot\sum_k\|a_k\|_2^2}{2\cdot\sum_k a_k}$$

$$\min_{i=1,\dots,k} E[f(x^{(i)})] \to f^*$$

Jensen's inequality and concavity of minimum yields:

$$E[f_{best}^{(k)}] = E\left[\min_{i=1,\dots,k} f(x^{(i)})\right] \le \min_{i=1,\dots,k} E[f(x^{(i)})]$$

$E[f_{best}^{(k)}] \to f^*$ convergence in expectation

Markov's inequality

$$\mathrm{Prob}\left(f_{best}^{(k)} - f^* \ge \epsilon\right) \le \frac{E[f_{best}^{(k)}-f^*]}{\epsilon}$$

rhs goes to 0, so we get convergence in probability eg. stochastic programming

$$\begin{aligned} \min\ & E f_0(x, w) \\ \text{s.t.}\ & E f_i(x, w) \le 0 \end{aligned}$$

(posterior analysis)

a convex problem is convex if $f_i(x, w)$ is convex in x for each w