Convex Optimization and Modeling

First Order Methods

11th lecture, 23.06.2010

Jun.-Prof. Matthias Hein

Practice: Problems are often too large to use second-order methods.

First Order Methods:

- Projection onto convex sets,
- Projected Subgradient, Projected Gradient, Projected Newton
- Proximal Gradient Methods (ISTA, FISTA) (next week)

References:

- Bertsekas: Nonlinear Programming
- Beck, Teboulle: "Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems", SIAM J. Imaging Sciences, 2:183-202,(2009).

Projection onto closed, convex set C

$$\min_{x \in C} \|z - x\|_2^2$$

Theorem:

- The minimum is unique and is denoted by $P_C(z)$
- $x \in C$ is equal to $P_C(z)$ iff

$$\langle y - x, z - x \rangle \leq 0, \quad \forall\, y \in C.$$

- The projection is continuous and non-expansive,

$$\|P_C(x) - P_C(y)\| \leq \|x - y\| .$$

Proof:

- Existence as in exercise - as $\|z - x\|^2$ is strictly convex in x we get uniqueness,
- First-order optimality condition,

$$\langle \nabla f(x^*), y - x^* \rangle \geq 0, \quad \forall\, y \in C$$

Using $\nabla f(x^*) = -2(z - x^*)$ we get the result.

- $P_C(y), P_C(z) \in C$,

$$\langle z - P_C(z), P_C(y) - P_C(z) \rangle \leq 0, \quad \langle y - P_C(y), P_C(z) - P_C(y) \rangle \leq 0.$$

Addition yields $\langle P_C(y) - P_C(z), z - P_C(z) - y + P_C(y) \rangle \leq 0$

$$\langle P_C(y) - P_C(z), y - z \rangle \geq \|P_C(y) - P_C(z)\|^2 .$$

Result follows using Cauchy-Schwarz for the left-hand side.

Examples of projections onto convex sets C

- Affine set $C = \{x \mid Ax = b\}$ with $A \in \mathbb{R}^{p \times n}$ and $\text{rank}(A) = p$,

$$P_C(z) = z + A^T(AA^T)^{-1}(b - Az).$$

  Computationally inexpensive if p is small or $AA^T = \mathbb{1}$.

- Euclidean-Ball, $C = \{x \mid \|x\|_2 \leq 1\}$, $P_C(z) = \frac{z}{\|z\|_2}$,

- Box, $C = [a, b]^n$,

$$(P_C(z))_i = \begin{cases} a_i & \text{if } z_i \leq a_i, \\ z_i & \text{if } a_i < z_i < b_i, \\ b_i & \text{if } z_i \geq b_i. \end{cases}$$

- Non-negative Orthant, $C = R_+^n$,

$$(P_C(z))_i = \max\{0, z_i\}.$$

Examples of projections onto convex sets C (continued)

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

$$(P_C(z))_i = \begin{cases} z_i - \lambda, & \text{if } z_i \geq \lambda, \\ 0, & \text{if } -\lambda < z_i < \lambda, \\ z_i + \lambda, & \text{if } z_i \leq -\lambda. \end{cases}$$

  where $\lambda = 0$ if $\|z\|_1 \leq 1$ and otherwise λ is the solution of

$$\sum_{i=1}^{n} \max\{|z_i| - \lambda, 0\} = 1.$$

- Positive semidefinite cone, $C = S_+^n$. Let $Z \in S^n$, $Z = \sum_{i=1}^n \lambda_i u_i u_i^T$, then

$$P_C(Z) = \sum_{i=1}^{n} \max\{0, \lambda_i\} u_i u_i^T,$$

Constrained Optimization:

$$\min_{x \in C} f(x)$$

- f is continuously differentiable and gradient has Lipschitz constant L,

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

- C is convex and closed
- efficient if projection P_C onto C can be easily computed

Main Ingredient:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2 .$$

Later: Projected Newton, Projected Subgradient.

Projected Gradient II

Quadratic upper bound at current point x^k: For $t \leq \frac{1}{L}$,

$$\begin{aligned} x' &= \arg\min_{x \in C} \; f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle + \frac{1}{2t} \left\| x - x^k \right\|^2 \\ &= \arg\min_{x \in C} \; \left\| x - \left(x^k - t \nabla f(x^k)\right) \right\|^2 \\ &= P_C\Big(x^k - t\,\nabla f(x^k)\Big) \end{aligned}$$

Next iterate:

$$x^{k+1} = P_C\Big(x^k - t^k\,\nabla f(x^k)\Big),$$

Variant:

$$x^{k+1} = x^k + \alpha^k \left(P_C\Big(x^k - t^k\,\nabla f(x^k)\Big) - x^k \right).$$

How to choose t ?

- **Constant Stepsize:** Given that an upper bound $M \geq L$ on the Lipschitz-constant is known set $t = \frac{1}{M}$.
- **Backtracking Linesearch:** Let $\beta \in (0, 1)$. Define

$$x^+(t) = P_C\Big(x^k - t\,\nabla f(x^k)\Big).$$

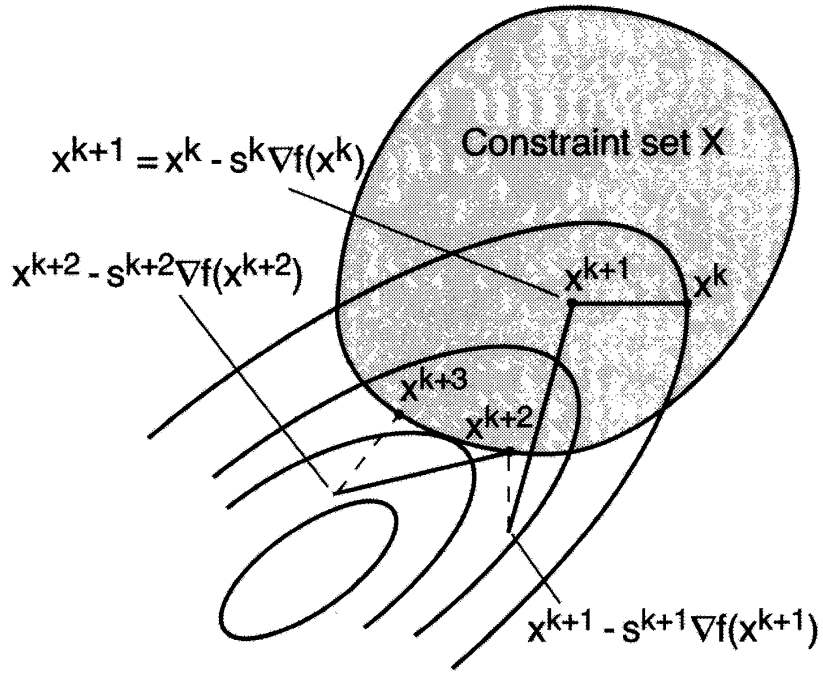
Start with $t_{\max}$ and then repeat $t := \beta t$ until,

$$f(x^+(t)) \leq f(x^k) + \Big\langle \nabla f(x^k), x^+(t) - x^k \Big\rangle + \frac{1}{2t}\left\| x^+(t) - x^k \right\|^2.$$

Basically, this is a test if $\frac{1}{t}$ fulfills the upper bound under which x' has been derived.

Projected Gradient III

Iteration: $x^{k+1} = P_C\Big(x^k - t^k \nabla f(x^k)\Big),$



If $x^k - t^k \nabla f(x^k) \in C$ we get a steepest descent step,

$$x^{k+1} = x^k - t^k \nabla f(x^k).$$

Optimality condition:

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0, \quad \forall\, x \in C.$$

For $t > 0$ this is equivalent to

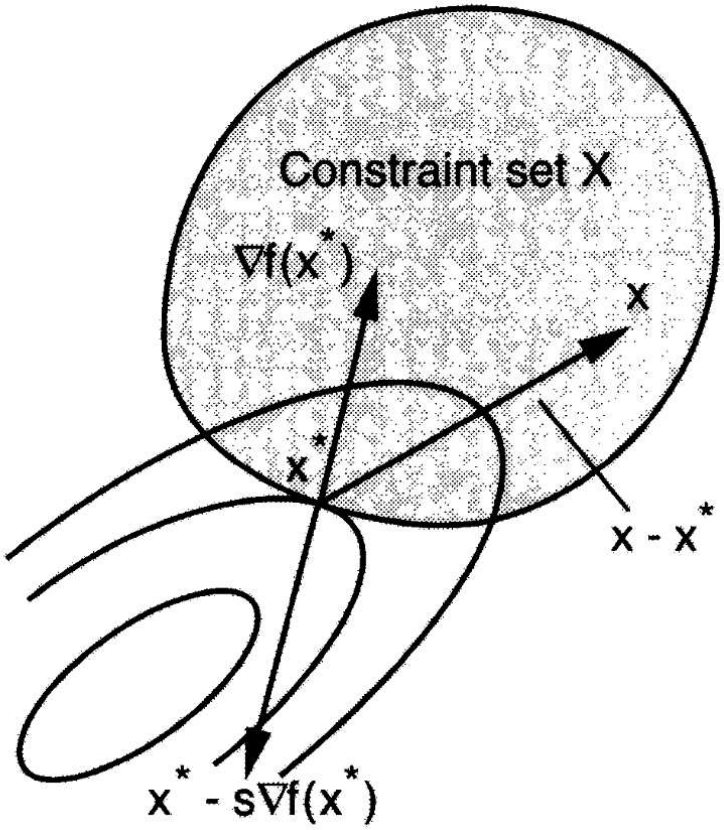
$$\left\langle \left(x^* - t\nabla f(x^*) \right) - x^*, x - x^* \right\rangle \leq 0, \quad \forall\, x \in C.$$

which is in turn equivalent to: $P_C\left(x^* - t\nabla f(x^*) \right) = x^*$.

Stopping criterion:

$$\left\| P_C\left(x^{(k)} - t^k \nabla f(x^{(k)}) \right) - x^{(k)} \right\| \leq \varepsilon\, t.$$

Optimality condition:



$$P_C\left(x^* - \frac{1}{L}\nabla f(x^*)\right) = x^*.$$

Non-increasing Sequence:

Theorem 1. *The iterates of the projected gradient method satsify*

$$f(x^{(k+1)}) \leq f(x^k).$$

Proof. Note, that with $x^k \in C$, we get

$$\begin{aligned} f(x^+(t)) &= \arg\min_{x \in C} \; f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle + \frac{1}{2t} \left\| x - x^k \right\|^2 \\ &\leq f(x^k) \end{aligned}$$

□

Gradient map:

$$G_t(x) = \frac{1}{t}\Big(x - P_C\big(x - t\nabla f(x)\big)\Big).$$

Note that,

$$P_C\big(x - t\nabla f(x)\big) = x - tG_t(x).$$

and from the optimality condition of the Projection P_C,

$$\big\langle x - t\nabla f(x) - P_C\big(x - t\nabla f(x)\big), z - P_C\big(x - t\nabla f(x)\big)\big\rangle \leq 0, \quad \forall\, z \in C.$$

one obtains

$$\langle G_t(x) - \nabla f(x), z - x + tG_t(x)\rangle \leq 0, \quad \forall\, z \in C.$$

Theorem 2. *The iterate $f(x^{(k)})$ of the projected gradient method satisfies*

- *for constant stepsize $0 < t \leq \frac{1}{L}$,*

$$f(x^{(k)}) - p^* \leq \frac{1}{2\,k\,t} \left\| x^{(0)} - x^* \right\|^2 .$$

- *for backtracking line-search,*

$$f(x^{(k)}) - p^* \leq \frac{1}{2\,k\,t_{\min}} \left\| x^{(0)} - x^* \right\|^2 ,$$

where $t_{\min} = \min_{i=1,\ldots,k} t^i$.

- much **slower** than the convergence rate of the gradient method for strictly convex functions with bounds on the Hessian,
- much **faster** than the $O(\frac{1}{\sqrt{k}})$ convergence rate of the subgradient method
- Projection does not influence the convergence rate.

Projected Gradient - Proof

Lemma: We have for all $y \in C$,

$$f(P_C\big(x - t\nabla f(x)\big)) \leq f(y) + \langle G_t(x), x - y\rangle - t(1 - \frac{tL}{2}) \left\|G_t(x)\right\|^2 .$$

Proof:

$$\begin{aligned}
f(x - tG_t(x)) &\leq f(x) - t\left\langle \nabla f(x), G_t(x)\right\rangle + \frac{L\,t^2}{2}\left\|G_t(x)\right\|^2 \\
&\leq f(x) - t\left\langle \nabla f(x) - G_t(x), G_t(x)\right\rangle - t(1 - \frac{t\,L}{2})\left\|G_t(x)\right\|^2 \\
&\leq f(y) + \left\langle \nabla f(x), x - y\right\rangle - t\left\langle \nabla f(x) - G_t(x), G_t(x)\right\rangle - t(1 - \frac{t\,L}{2})\left\|G_t(x)\right\|^2 \\
&\leq f(y) - \left\langle \nabla f(x) - G_t(x), y - x + tG_t(X)\right\rangle + \left\langle G_t(x), x - y\right\rangle - t(1 - \frac{t\,L}{2})\left\|G_t(x)\right\|^2 \\
&\leq f(y) + \left\langle G_t(x), x - y\right\rangle - t(1 - \frac{t\,L}{2})\left\|G_t(x)\right\|^2 ,
\end{aligned}$$

where the last equality follows from the optimality condition of the projection P_C.

Projected Gradient - Proof

Lemma: We have for all $y \in C$,

$$f(P_C\big(x - t\nabla f(x)\big)) \leq f(y) + \langle G_t(x), x - y\rangle - t(1 - \frac{tL}{2}) \left\|G_t(x)\right\|^2 .$$

Proof:

$$\begin{aligned}
f(x - tG_t(x)) &\leq f(x) - t\left\langle \nabla f(x), G_t(x)\right\rangle + \frac{L\,t^2}{2}\left\|G_t(x)\right\|^2 \\
&\leq f(x) - t\left\langle \nabla f(x) - G_t(x), G_t(x)\right\rangle - t(1 - \frac{t\,L}{2})\left\|G_t(x)\right\|^2 \\
&\leq f(y) + \left\langle \nabla f(x), x - y\right\rangle - t\left\langle \nabla f(x) - G_t(x), G_t(x)\right\rangle - t(1 - \frac{t\,L}{2})\left\|G_t(x)\right\|^2 \\
&\leq f(y) - \left\langle \nabla f(x) - G_t(x), y - x + tG_t(X)\right\rangle + \left\langle G_t(x), x - y\right\rangle - t(1 - \frac{t\,L}{2})\left\|G_t(x)\right\|^2 \\
&\leq f(y) + \left\langle G_t(x), x - y\right\rangle - t(1 - \frac{t\,L}{2})\left\|G_t(x)\right\|^2 ,
\end{aligned}$$

where the last equality follows from the optimality condition of the projection P_C.

Proof of Theorem: We have for $0 < t < \frac{1}{L}$,

$$\begin{aligned} f(P_C\big(x - t\nabla f(x)\big)) &\leq f(x^*) + \langle G_t(x), x - x^* \rangle - t(1 - \frac{tL}{2}) \left\| G_t(x) \right\|^2 \\ &\leq f(x^*) + \langle G_t(x), x - x^* \rangle - \frac{t}{2} \left\| G_t(x) \right\|^2 \\ &\leq f(x^*) + \frac{1}{2t}\Big(\left\| x - x^* \right\|^2 - \left\| x - x^* - tG_t(x) \right\|^2 \Big) \end{aligned}$$

Note, that this implies: $\left\| x^{(k)} - x^* \right\| \geq \left\| x^{(k+1)} - x^* \right\|$. Using that $f(x^{(k)})$ is a non-increasing sequence, we get

$$\begin{aligned} f(x^{(k)}) - f(x^*) &\leq \frac{1}{k} \sum_{i=0}^{k-1} \big(f(x^{(i)}) - f(x^*) \big) \\ &\leq \frac{1}{2t_{\min} k}\Big(\left\| x^{(0)} - x^* \right\|^2 - \left\| x^{(k)} - x^* \right\|^2 \Big) \end{aligned}$$

Assumption: f is twice-differentiable.

Quadratic Taylor expansion at current point x^k

$$x' = \arg\min_{x \in C} \; f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle + \frac{1}{2} \left\langle x - x^k, Hf(x^k)(x - x^k) \right\rangle$$
$$= P_{(Hf)^{\frac{1}{2}} C} \left((Hf)^{\frac{1}{2}} \left(x^k - (Hf(x^k))^{-1} \nabla f(x^k) \right) \right)$$

$\Longrightarrow x' \in C$ is the closest point to $x^k - (Hf(x^k))^{-1} \nabla f(x^k)$ but with respect to the norm $\|z\|_{Hf} = \sqrt{\langle z, Hf(x^k) z \rangle}$.

Next iterate: with stepsize parameter $\alpha^k \in (0, 1]$

$$x^{k+1} = x^k + \alpha^k (x' - x^k),$$

- x^{k+1} convex combination of x^k and x' and thus in C,

Problem: $\min_{x \in \mathbb{R}^n} f(x)$ - no assumptions on f.

Subgradient method:

$$x^{k+1} = x^k - \alpha^k g^k, \quad \alpha^k > 0, \; g_k \in \partial f(x^k).$$

- No stepsize selection ! Sequence α^k will be fixed initially.
- **Any** subgradient is o.k. ! Do not have to know $\partial f(x^k)$ - one subgradient for each point is sufficient.

Projected Subgradient method:

$$x^{k+1} = P_C\big(x^k - \alpha^k\, g^k\big), \quad \alpha^k > 0,\ g_k \in \partial f(x^k).$$

- convergence analysis applies almost without change, for any $y \in C$,

$$\begin{aligned}\left\|x^{k+1} - y\right\|^2 &= \left\|P_C(x^k - \alpha^k\, g^k) - y\right\|^2 \\ &= \left\|P_C(x^k - \alpha^k\, g^k) - P_C(y)\right\|^2 \leq \left\|x^k - \alpha^k\, g^k - y\right\|^2\end{aligned}$$

- convergence speed is $O(\frac{1}{\sqrt{k}})$ as for the unconstrained problem (for suitable choice of the sequence α^k).

Non-negative L_1

$$\min \ \|Ax - b\|_1$$
$$\text{subject to: } x \in \mathbb{R}^n_+$$

The iterate becomes,

$$x^{k+1} = (x^k - \alpha^k A^T \operatorname{sign}(Ax - b))_+.$$

Example: Projected Subgradient II

Non-negative L_1: with $\alpha^k = \frac{1}{10k}$,

$$x^{k+1} = (x^k - \alpha^k A^T \operatorname{sign}(Ax - b))_+.$$