Agenda

Proximal gradient methods

1. Gradient descent
2. Proximal gradient method
3. Prox functions
4. Subgradients
5. Convergence of proximal gradient methods

Classical gradient descent

$$\min \quad f(x)$$

- f cvx
- f differentiable

Choose x_0 and repeat

$$x_k = x_{k-1} - t_k \nabla f(x_{k-1}) \quad k = 1, 2, \ldots$$

Step size rules

- $t_k = \hat{t}$ cst
- Backtracking line search [cf. Boyd and Vandenberghe]
- Exact line search
$$\min_t \quad f(x - t\nabla f(x))$$
- Barzila-Borwein
- ...

Convergence of gradient descent

- $\text{dom}(f) = \mathbb{R}^n$
- minimizer $x^\star$ with optimal value $f^\star = f(x^\star)$
- ∇f is Lipschitz continuous with $L > 0$

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

(or $\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|$ where $\|.\|$ and $\|\cdot\|_*$ are dual from each other)
If f is twice differentiable, this means $\nabla^2 f \preceq L\, I$

Convergence

Fix step size $t \leq 1/L$. Then

$$f(x_k) - f^\star \leq \frac{\|x_0 - x^\star\|^2}{2tk}$$

Convergence of gradient descent

With backtracking line search

- Initialize at $\hat{t} > 0$
- Take $t = \beta t$ until

$$f(x_k - t\nabla f(x_k)) \leq f(x_k) - \alpha t \|\nabla f(x_k)\|^2$$

where $0 < \alpha, \beta < 1$; e. g. $\alpha = 0.5$

Convergence

Set $t_{\min} = \min\{\hat{t}, \frac{\beta}{L}\}$, then

$$f(x_k) - f^\star \leq \frac{\|x_0 - x^\star\|^2}{2 t_{\min} k}$$

Interpretation of gradient descent

$$
\begin{aligned}
x &= x_0 - t\nabla f(x_0) \\
&= \text{arg min}_x \left\{ f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{1}{2t}\|x - x_0\|^2 \right\}
\end{aligned}
$$

$f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle$	first-order approximation to objective
$\frac{1}{2t}\|x - x_0\|^2$	proximity term with weight $\frac{1}{2t}$

Composite functions

$$\min \quad f(x) = g(x) + h(x)$$

- g cvx and diff
- h cvx (not necessarily diff)

Examples:

1. $h = 0 \quad \rightarrow \quad \min \quad g(x)$
2. $h = \mathbf{1}_C$ 'indicator' of cvx set: $h(x) = 0$ if $x \in C$ and ∞ otherwise

$$\min\ f(x) \quad \Leftrightarrow \quad \begin{array}{ll} \min & g(x) \\ \text{s.t.} & x \in C \end{array}$$

3. $h(x) = \|x\|_1$

$$\min \quad g(x) + \|x\|_1$$

e.g. $g(x) = \frac{1}{2\lambda}\|Ax - b\|_2^2$
g cvx and diff; $\nabla^2 g = \frac{1}{\lambda}A^*A$ and ∇g Lipschitz with $L = \frac{\|A\|^2}{\lambda}$

Generalized gradient step

$$x = \text{argmin}\{g(x_0) + \langle \nabla g(x_0), x - x_0 \rangle + \frac{1}{2t}\|x - x_0\|^2 + h(x)\}$$

- quadratic approximation to g only
- computation of gradient step $\rightarrow$ later

Examples:

- $h = 0 \quad \rightarrow \quad$ gradient step
- $h = \mathbf{1}_C$

$$\begin{aligned} x &= \text{argmin}\left\{\frac{1}{2t}\|x - (x_0 - t\nabla g(x_0))\|^2 + h(x)\right\} \\ &= \text{argmin}_{x \in C}\left\{\frac{1}{2t}\|x - (x_0 - t\nabla g(x_0))\|^2\right\} \end{aligned}$$

$P_c =$ Projection onto cvx set C

$$x = P_c(x_0 - t\nabla g(x_0))$$

this is a projected gradient step

- $h(x) = \|x\|_1$

$$x = \text{arg min}\left\{\frac{1}{2t}\|x - (x_0 - t\nabla g(x_0))\|^2 + \|x\|_1\right\}$$
$$= S_t(x_0 - t\nabla g(x_0))$$

S_t: Soft-thresholding/shrinkage operator

$$S_t(z) = \text{arg min}_x \quad \frac{1}{2t}[\|z - x\|_2^2 + \|x\|_1]$$
$$= \text{arg min}_x \quad \sum_i \frac{1}{2t}(z_i - x_i)^2 + |x_i|$$

$$[S_t(z)]_i = \begin{cases} z_i - t & z_i \geq t \\ 0 & |z_i| \leq t \\ z_i + t & z_i \leq -t \end{cases}$$

Proximation gradient method I

$$\min \quad g(x) + h(x)$$

- Proximal operator

$$\text{prox}_{th}(z) = \text{argmin}_x \left\{ \frac{1}{2t} \|x - z\|^2 + h(x) \right\}$$

prox well defined if $t > 0$ and $\text{dom}(h) = \mathbb{R}^n$ (unique minimizer for all z)

- Proximal step

$$\begin{aligned} x &= \text{argmin} \left\{ \frac{1}{2t} \|x - [x_0 - t\nabla g(x_0)]\|^2 + h(x) \right\} \\ &= \text{prox}_{th}(x_0 - t\nabla g(x_0)) \end{aligned}$$

- Remarks
  (i) $h = \mathbf{1}_C \Rightarrow S_t = P_C$
  (ii) $h = 0 \Rightarrow x = x_0 - t\nabla g(x_0)$

Proximal gradient method II

$$\text{min} \quad g(x) + h(x)$$

Choose x_0 and repeat for $k = 1, 2, \ldots$

$$x_k = \text{prox}_{t_k h}(x_{x-1} - t_k \nabla g(x_{k-1}))$$

Can be applied with fixed step sizes/backtracking line search

Example: Lasso

$$\text{min} \quad \frac{1}{2\lambda}\|Ax - b\|^2 + \|x\|_1$$

Choose x_0 and repeat

$$x_k = \text{shrink}(x_{k-1} - \lambda^{-1} t_k A^*(Ax_{k-1} - b); t_k)$$

called Iterated Soft-Thresholding Algorithm (ISTA)

Subgradients

- f cvx
- v is a subgradient of f at x_0 denoted by $v \in \partial f(x_0)$
  If for all $x \in$ dom(f)

$$f(x) \geq f(x_0) + v^T(x - x_0)$$

  Subdifferential $\partial f(x_0)$: set of all subgradients

$$\partial f(x_0) \neq \emptyset$$

- $x^\star$ minimizes $f(x)$ iff $0 \in \partial f(x^\star)$
- Remark: f diff $\Rightarrow \partial f(x) = \{\nabla f(x)\}$ and optimality condition is $0 = \nabla f(x^\star)$

Optimality conditions

$$\min \quad f(x) = g(x) + h(x)$$

$$x \text{ optimal} \quad \Leftrightarrow \quad \nabla g(x) + v = 0 \text{ and } v \in \partial h(x)$$

Proposition

x optimal iff $x = \mathrm{prox}_{th}(x - t\nabla g(x))$ for any $t > 0$. That is, iff x is fixed point of update rule

Proof:

$$\begin{aligned} \text{Fixed point} &\Leftrightarrow x \text{ is a minimizer of } z \mapsto \frac{1}{2t}\|z - (x - t\nabla g(x))\|^2 + h(z) \\ &\Leftrightarrow \nabla g(x) + v = 0 \text{ and } v \in \partial h(x) \end{aligned}$$

Monotonicity I

$$x_+ = \text{arg min } \{\frac{1}{2t}\|z - (x - t\nabla g(x))\|^2 + h(z)\} \triangleq x - tG_t(x)$$

Optimality condition:

$$v = G_t(x) - \nabla g(x) \text{ and } v \in \partial h(x_+)$$

Hence,

$$h(x_+) \leq h(y) + \langle v, x_+ - y \rangle$$

Monotonicity II

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

This gives

$$\begin{aligned} f(x_+) &\leq g(x) - t\langle \nabla g(x), G_t(x) \rangle + \frac{Lt^2}{2}\|G_t(x)\|^2 + h(x_+) \\ &= g(x) + t\langle v, G_t(x) \rangle - t\Big(1 - \frac{Lt}{2}\Big)\|G_t(x)\|^2 + h(x_+) \\ &\leq g(x) + t\langle v, G_t(x) \rangle - t\Big(1 - \frac{Lt}{2}\Big)\|G_t(x)\|^2 + h(y) + \langle v, x_+ - y \rangle \\ &= g(x) - t\Big(1 - \frac{Lt}{2}\Big)\|G_t(x)\|^2 + h(y) + \langle v, x - y \rangle \end{aligned}$$

Since $g(x) \leq g(y) + \langle \nabla g(x), x - y \rangle$

$$f(x_+) \leq f(y) + \langle G_t(x), x - y \rangle - t\Big(1 - \frac{Lt^2}{2}\Big)\|G_t(x)\|^2 \quad \forall x, y$$

$$x = y \quad \implies f(x_+) \le f(x) - t\Big(1 - \frac{Lt}{2}\Big)\|G_t(x)\|^2$$

Conclusion: If $t < \frac{2}{L}$, each step decreases objective function value unless $G_t(x) = 0$. But then we're at optimum!

Convergence of proximal gradient method

- f has an optimal solution $x^\star$ and $f^\star = f(x^\star)$
- g and h cvx
- ∇g Lipschitz with cst $L > 0$

Theorem:

Fix step size $t \leq \frac{1}{L}$

$$f(x_k) - f^\star \leq \frac{\|x_0 - x^\star\|^2}{2tk}$$

Similar with backtracking $t \leftarrow$ *min*$(\hat{t}, \frac{\beta}{L})$

Proof

$0 \leq t \leq 1/L$

$$\begin{aligned} f(x_k) &\leq f^\star + G_t(x_{k-1})^T(x_{k-1} - x^\star) - \frac{t}{2}\|G_t(x_{k-1})\|^2 \\ &= f^\star + \frac{1}{2t}\left[\|x_{k-1} - x^\star\|^2 - \|x_{k-1} - tG_t(x_{k-1}) - x^\star\|^2\right] \\ &= f^\star + \frac{1}{2t}\left[\|x_{k-1} - x^\star\|^2 - \|x_k - x^\star\|^2\right] \end{aligned}$$

This implies

$$k[f(x_k) - f^\star] \leq \sum_{j=1}^{k} [f(x_j) - f^\star] \leq \frac{1}{2t}\|x_0 - x^\star\|^2 \quad \Rightarrow \quad f(x_k) - f^\star \leq \frac{\|x_0 - x^\star\|^2}{2tk}$$

Same analysis with backtracking because $\alpha \geq 1/2$

$$t_k \geq t_{\mathsf{min}} = \mathsf{min}\{\hat{t}, \beta/L\}$$

Therefore

$$f(x_k) - f^\star \leq \frac{\|x_0 - x^\star\|^2}{2t_{\mathsf{min}}k}$$

Main pillar of analysis

We have established this useful result:

- Fix $t > 0$ and set $x_+ = x - tG_t(x)$
- Assume $f(z) \leq Q_{1/t}(z, x)$

  Then $\forall y \in \mathrm{dom}(f)$

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

Philosophy: majorization–minimization

1. Find "relevant" approximation to objective such that

   (i) $f(x) = \mu(x,x) \quad \forall x$

   (ii) $f(x) \leq \mu(x,y) \quad \forall x, y$

2. Minimization scheme

$$x_k = \text{argmin}_x \mu(x, x_{k-1})$$

   which implies $\mu(x_k, x_{k-1}) \leq \mu(x, x_{k-1})\ \forall x$

$$f(x_k) \leq \mu(x_k, x_{k-1}) \leq \mu(x_{k-1}, x_{k-1}) = f(x_{k-1})$$

   $\implies$ minimizing sequence

Question: How to generate a good upper bound?

Generalized gradient descent operates by upper bounding smooth component by simple quadratic term

$$\mu(x, y) = g(y) + \langle \nabla g(y), x - y \rangle + \frac{1}{2t} \|x - y\|^2 + h(x)$$

Special cases

1. $h = 0 \rightarrow$ convergence of gradient descent
2. $h = \mathbf{1}_C \rightarrow$ convergence of projected gradient descent
3. $g = 0 \rightarrow$ this is proximal minimization

Proximal minimization algorithm (PMA)

$$\text{min} \quad h(x)$$

Generalized GD reduces to PMA. Choose x_0 and for $k = 1, 2, \ldots$

$$x_k = \text{argmin} \left\{ \frac{1}{2t_k} \|x - x_{k-1}\|^2 + h(x) \right\}$$

Theorem

Set $\sigma_k = \sum_{j \le k} t_j$

$$h(x_k) - h^\star \le \frac{\|x - x^\star\|^2}{2\sigma_k}$$

- Algorithm is better than subgradient methods but not implementable unless h is ‘simple’
- Very useful when combined with duality $\longrightarrow$ augmented Lagrangian methods

Subgradient methods

$$\left| \begin{array}{ll} \text{min} & h(x) \\ \text{subject to} & x \in C \end{array} \right.$$

Subgradient scheme:

$$v_{k-1} \in \partial h(x_{k-1}) \qquad \text{and} \qquad x_k = P_C(x_{k-1} - t_k v_{k-1})$$

Scheme is not monotone

Typical result

- h cvx and Lipschitz, $\|h(x) - h(y)\| \leq \mu\|x - y\|$
- C cvx and compact
- Set $t_k = \frac{\text{diam}(C)}{\sqrt{k}}$

Then

$$\min_{1 \leq j \leq k} h(x_j) - h^\star \leq O(\mu)\,\frac{\text{diam}(C)}{\sqrt{k}}$$

Summary

- Generalized GD $\rightarrow$ convergence rate $\frac{1}{k}$
- Subgradient methods $\rightarrow$ convergence rate $\frac{1}{\sqrt{k}}$

Can we do better for non-smooth problems

$$\min\ f(x) = g(x) + h(x)$$

with the same computational effort as generalized GD but with faster convergence?

Answer: Yes we can - with equally simple scheme

$$x_{k+1} = \arg\min\ Q_{1/t}(x, y_k)$$

Note that we use y_k instead of x_k where new point is cleverly chosen

- Original idea: Nesterov 1983 for minimization of smooth objective
- Here: nonsmooth problem

References

1. Y. Nesterov. *Gradient methods for minimizing composite objective function* Technical Report – CORE – Université Catholique de Louvain, (2007)

2. A. Beck and M. Teboulle. Fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM J. Imaging Sciences*, (2008)

3. M. Teboulle, First Order Algorithms for Convex Minimization, Optimization Tutorials (2010), IPAM, UCLA

4. L. Vandenberghe, EE236C (Spring 2011), UCLA