Proximal Gradient Method

Pontus Giselsson

Outline

- **A fundamental inequality**
- Nonconvex setting
- Convex setting
- Strongly convex setting
- Backtracking
- Stopping conditions
- Accelerated gradient method
- Scaling

Proximal gradient method

- We consider composite optimization problems of the form

$$\underset{x}{\text{minimize}}\, f(x) + g(x)$$

- The proximal gradient method is

$$\begin{aligned}
x_{k+1} &= \underset{y}{\text{argmin}} \Big(f(x_k) + \nabla f(x_k)^T (y - x_k) + \tfrac{1}{2\gamma_k} \|y - x_k\|_2^2 + g(y) \Big) \\
&= \underset{y}{\text{argmin}} \Big(g(y) + \tfrac{1}{2\gamma_k} \|y - (x_k - \gamma_k \nabla f(x_k))\|_2^2 \Big) \\
&= \text{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k))
\end{aligned}$$

Proximal gradient – Optimality condition

- Proximal gradient iteration is:

$$\begin{aligned} x_{k+1} &= \mathrm{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k)) \\ &= \underset{y}{\mathrm{argmin}}(g(y) + \underbrace{\tfrac{1}{2\gamma_k}\|y - (x_k - \gamma_k \nabla f(x_k))\|_2^2}_{h(y)}) \end{aligned}$$

  where x_{k+1} is unique due to strong convexity of h

- Fermat's rule gives, since g convex, optimality condition:

$$\begin{aligned} 0 &\in \partial g(x_{k+1}) + \partial h(x_{k+1}) \\ &= \partial g(x_{k+1}) + \gamma_k^{-1}(x_{k+1} - (x_k - \gamma_k \nabla f(x_k))) \end{aligned}$$

  since h differentiable

- A consequence is that $\partial g(x_{k+1})$ is nonempty

Proximal gradient method – Convergence rates

- We will analyze proximal gradient method in different settings:
  - Nonconvex
    - $O(1/k)$ convergence for squared residual
  - Convex
    - $O(1/k)$ convergence for function values
  - Strongly convex
    - Linear convergence in distance to solution
- First two rates based on a *fundamental inequality* for the method

Assumptions for fundamental inequality

(i) $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable (not necessarily convex)

(ii) For every x_k and x_{k+1} there exists $\beta_k \in [\eta, \eta^{-1}]$, $\eta \in (0, 1]$:

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \tfrac{\beta_k}{2} \|x_k - x_{k+1}\|_2^2$$

where β_k is a sort of local Lipschitz constant

(iii) $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is closed convex

(iv) A minimizer $x^\star$ exists and $p^\star = f(x^\star) + g(x^\star)$ is optimal value

(v) Proximal gradient method parameters $\gamma_k > 0$

- Assumption (ii) satisfied with $\beta_k \geq \beta$ if f is β-smooth
- Assumptions will be strengthened later

A fundamental inequality

For all $z \in \mathbb{R}^n$, the proximal gradient method satisfies

$$
\begin{aligned}
f(x_{k+1}) + g(x_{k+1}) \le f(x_k) + \nabla f(x_k)^T(z - x_k) - \tfrac{\gamma_k^{-1} - \beta_k}{2}\|x_{k+1} - x_k\|_2^2 \\
+ g(z) + \tfrac{1}{2\gamma_k}(\|x_k - z\|_2^2 - \|x_{k+1} - z\|_2^2)
\end{aligned}
$$

where $x_{k+1} = \mathrm{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k))$

A fundamental inequality – Proof (1/2)

Using

(a) Upper bound assumption on f, i.e., Assumption (ii)

(b) Prox optimality condition: There exists $s_{k+1} \in \partial g(x_{k+1})$

$$0 = s_{k+1} + \gamma_k^{-1}(x_{k+1} - (x_k - \gamma_k \nabla f(x_k)))$$

(c) Subgradient definition: $\forall z, g(z) \geq g(x_{k+1}) + s_{k+1}^T(z - x_{k+1})$

$$\begin{aligned}
&f(x_{k+1}) + g(x_{k+1}) \\
&\qquad \overset{(a)}{\leq} f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \tfrac{\beta_k}{2}\|x_{k+1} - x_k\|_2^2 + g(x_{k+1}) \\
&\qquad \overset{(c)}{\leq} f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \tfrac{\beta_k}{2}\|x_{k+1} - x_k\|_2^2 + g(z) \\
&\qquad\quad - s_{k+1}^T(z - x_{k+1}) \\
&\qquad \overset{(b)}{=} f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \tfrac{\beta_k}{2}\|x_{k+1} - x_k\|_2^2 + g(z) \\
&\qquad\quad + \gamma_k^{-1}(x_{k+1} - (x_k - \gamma_k \nabla f(x_k)))^T(z - x_{k+1}) \\
&\qquad = f(x_k) + \nabla f(x_k)^T(z - x_k) + \tfrac{\beta_k}{2}\|x_{k+1} - x_k\|_2^2 + g(z) \\
&\qquad\quad + \gamma_k^{-1}(x_{k+1} - x_k)^T(z - x_{k+1})
\end{aligned}$$

A fundamental inequality – Proof (2/2)

- The proof continues by using the equality

$$\begin{aligned}&(x_{k+1} - x_k)^T(z - x_{k+1})\\ &\qquad\qquad = \tfrac{1}{2}(\|x_k - z\|_2^2 - \|x_{k+1} - z\|_2^2 - \|x_{k+1} - x_k\|_2^2)\end{aligned}$$

- Applying to previous inequality gives

$$\begin{aligned}&f(x_{k+1}) + g(x_{k+1})\\ &\qquad \le f(x_k) + \nabla f(x_k)^T(z - x_k) + \tfrac{\beta_k}{2}\|x_{k+1} - x_k\|_2^2 + g(z)\\ &\qquad\quad + \gamma_k^{-1}(x_{k+1} - x_k)^T(z - x_{k+1})\\ &\qquad = f(x_k) + \nabla f(x_k)^T(z - x_k) + \tfrac{\beta_k}{2}\|x_{k+1} - x_k\|_2^2 + g(z)\\ &\qquad\quad + \tfrac{1}{2\gamma_k}(\|x_k - z\|_2^2 - \|x_{k+1} - z\|_2^2 - \|x_k - x_{k+1}\|_2^2)\end{aligned}$$

which after rearrangement gives the fundamental inequality

Outline

- A fundamental inequality
- **Nonconvex setting**
- Convex setting
- Strongly convex setting
- Backtracking
- Stopping conditions
- Accelerated gradient method
- Scaling

Nonconvex setting

- We will analyze the proximal gradient method

$$x_{k+1} = \mathrm{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k))$$

in a nonconvex setting for solving

$$\mathrm{minimize}\, f(x) + g(x)$$

- Will show sublinear $O(1/k)$ convergence
- Analysis based on *A fundamental inequality*

Nonconvex setting – Assumptions

(i) $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable (not necessarily convex)

(ii) For every x_k and x_{k+1} there exists $\beta_k \in [\eta, \eta^{-1}]$, $\eta \in (0, 1]$:

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \tfrac{\beta_k}{2}\|x_k - x_{k+1}\|_2^2$$

where β_k is a sort of local Lipschitz constant

(iii) $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is closed convex

(iv) A minimizer $x^\star$ exists and $p^\star = f(x^\star) + g(x^\star)$ is optimal value

(v) Algorithm parameters $\gamma_k \in [\epsilon, \frac{2}{\beta_k} - \epsilon]$, where $\epsilon > 0$

- Differs from assumptions for fundamental inequality only in (v)
- Assumption (ii) satisfied with $\beta_k \geq \beta$ if f is β-smooth

Nonconvex setting – Analysis

- Use fundamental inequality

$$f(x_{k+1}) + g(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T(z - x_k) - \tfrac{\gamma_k^{-1} - \beta_k}{2}\|x_{k+1} - x_k\|_2^2$$
$$+ g(z) + \tfrac{1}{2\gamma_k}(\|x_k - z\|_2^2 - \|x_{k+1} - z\|_2^2)$$

- Set $z = x_k$ to get

$$f(x_{k+1}) + g(x_{k+1}) \leq f(x_k) + g(x_k) - (\gamma_k^{-1} - \tfrac{\beta_k}{2})\|x_{k+1} - x_k\|_2^2$$

Step-size requirements

- Step-sizes γ_k should be restricted for inequality to be useful:

$$f(x_{k+1}) + g(x_{k+1}) \leq f(x_k) + g(x_k) - (\gamma_k^{-1} - \tfrac{\beta_k}{2})\|x_{k+1} - x_k\|_2^2$$

- Requirements $\beta_k \in [\eta, \eta^{-1}]$ and $\gamma_k \in [\epsilon, \frac{2}{\beta_k} - \epsilon]$:
  - upper bound $\gamma_k \leq \frac{2}{\beta_k} - \epsilon$ can be written as

$$\gamma_k \leq \tfrac{2}{\beta_k + 2\delta_k} \qquad \text{where} \qquad \delta_k = \tfrac{\beta_k \epsilon}{2\left(\frac{2}{\beta_k} - \epsilon\right)} \geq \tfrac{\beta_k^2 \epsilon}{4} \geq \tfrac{\eta^2 \epsilon}{4} > 0$$

    since upper bound $\beta_k \leq \eta^{-1}$ gives $\frac{2}{\beta_k} - \epsilon \geq 2\eta - \epsilon > 0$ and $\epsilon > 0$

  - Inverting upper step-size bound and letting $\delta := \frac{\eta^2 \epsilon}{4} \leq \delta_k$:

$$\gamma_k^{-1} \geq \tfrac{\beta_k + 2\delta_k}{2} \geq \tfrac{\beta_k}{2} + \delta \qquad \Rightarrow \qquad \gamma_k^{-1} - \tfrac{\beta_k}{2} \geq \delta > 0$$

- This implies, by subtracting $p^\star$ from both sides to have $V_k \geq 0$,

$$\underbrace{f(x_{k+1}) + g(x_{k+1}) - p^\star}_{V_{k+1}} \leq \underbrace{f(x_k) + g(x_k) - p^\star}_{V_k} - \underbrace{\delta\|x_{k+1} - x_k\|_2^2}_{R_k}$$

  where bounds on γ_k imply that all R_k are nonnegative

Lyapunov inequality consequences

- Restating Lyapunov inequality

$$\underbrace{f(x_{k+1}) + g(x_{k+1}) - p^\star}_{V_{k+1}} \leq \underbrace{f(x_k) + g(x_k) - p^\star}_{V_k} - \underbrace{\delta\|x_{k+1} - x_k\|_2^2}_{R_k}$$

- Consequences:
  - Function value is decreasing sequence (may not converge to $p^\star$)
  - Fixed-point residual converges to 0 as $k \to \infty$:

$$\|x_{k+1} - x_k\|_2 = \|\mathrm{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k)) - x_k\|_2 \to 0$$

  - Best fixed-point residual norm square converges as $O(1/k)$:

$$\min_{i \in \{0,\ldots,k\}} \|x_{i+1} - x_i\|_2^2 \leq \frac{f(x_0) + g(x_0) - p^\star}{\delta(k+1)}$$

Lyapunov inequality consequences – $g = 0$

- For $g = 0$, then $x_{k+1} = x_k - \gamma_k \nabla f(x_k)$ and

$$\|x_{k+1} - x_k\|_2 = \gamma_k \|\nabla f(x_k)\|_2 \qquad \text{and} \qquad R_k = \delta\gamma_k^2 \|\nabla f(x_k)\|_2^2$$

- Lyapunov inequality consequences in this setting:
  - Gradient converges to 0 (since $\gamma_k \geq \epsilon$): $\|\nabla f(x_k)\|_2 \to 0$
  - Smallest gradient norm square converges as:

$$\min_{i \in \{0,\ldots,k\}} \|\nabla f(x_i)\|_2^2 \leq \frac{f(x_0) - p^\star}{\delta \sum_{i=0}^k \gamma_i^2}$$

  - If, in addition, f is β-smooth and $\gamma_k = \frac{1}{\beta}$:

$$\min_{i \in \{0,\ldots,k\}} \|\nabla f(x_i)\|_2^2 \leq \frac{2\beta(f(x_0) - p^\star)}{k+1}$$

    since then $\beta_k = \beta$ and $\gamma_k^{-1} - \frac{\beta_k}{2} = \frac{\beta}{2} = \delta > 0$

- So, will approach local maximum, minimum, or saddle-point

Fixed-point residual convergence – Implication

What does $\|\text{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k)) - x_k\|_2 \to 0$ imply?

- By prox-grad optimality condition and $\|x_{k+1} - x_k\|_2 \to 0$:

  $$\partial g(x_{k+1}) + \nabla f(x_k) \ni \gamma_k^{-1}(x_k - x_{k+1}) \to 0$$

  as $k \to \infty$ (since $\gamma_k \geq \epsilon$, i.e., $0 < \gamma_k^{-1} \leq \epsilon^{-1}$) or equivalently

  $$\partial g(x_{k+1}) + \nabla f(x_{k+1}) \ni \underbrace{\gamma_k^{-1}(x_k - x_{k+1}) + \nabla f(x_{k+1}) - \nabla f(x_k)}_{u_k} \to 0$$

  where $u_k \to 0$ is concluded by continuity of ∇f
- Critical point definition for nonconvex f satisfied in the limit

Outline

- A fundamental inequality
- Nonconvex setting
- **Convex setting**
- Strongly convex setting
- Backtracking
- Stopping conditions
- Accelerated gradient method
- Scaling

Convex setting

- We will analyze the proximal gradient method

$$x_{k+1} = \operatorname{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k))$$

in the convex setting for solving

$$\operatorname{minimize} f(x) + g(x)$$

- Will show sublinear $O(1/k)$ convergence for function values
- Analysis based on *A fundamental inequality*

Convex setting – Assumptions

(i) $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and convex

(ii) For every x_k and x_{k+1} there exists $\beta_k \in [\eta, \eta^{-1}]$, $\eta \in (0, 1]$:

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \tfrac{\beta_k}{2} \|x_k - x_{k+1}\|_2^2$$

where β_k is a sort of local Lipschitz constant

(iii) $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is closed convex

(iv) A minimizer $x^\star$ exists and $p^\star = f(x^\star) + g(x^\star)$ is optimal value

(v) Algorithm parameters $\gamma_k \in [\epsilon, \frac{2}{\beta_k} - \epsilon]$, where $\epsilon > 0$

- Assumptions as for fundamental inequality plus
  - convexity of f
  - restricted step-size parameters γ_k (as in nonconvex setting)
- Assumption (ii) satisfied with $\beta_k \geq \beta$ if f is β-smooth

Convex setting – Analysis

- Use fundamental inequality with $z = x^\star$, where $x^\star$ is solution

$$\begin{aligned} f(x_{k+1}) + g(x_{k+1}) \leq f(x_k) &+ \nabla f(x_k)^T(x^\star - x_k) \\ &- \tfrac{\gamma_k^{-1} - \beta_k}{2}\|x_{k+1} - x_k\|_2^2 + g(x^\star) \\ &+ \tfrac{1}{2\gamma_k}(\|x_k - x^\star\|_2^2 - \|x_{k+1} - x^\star\|_2^2) \end{aligned}$$

- and convexity of f

$$f(x^\star) \geq f(x_k) + \nabla f(x_k)^T(x^\star - x_k)$$

- This gives

$$\begin{aligned} f(x_{k+1}) + g(x_{k+1}) \leq f(x^\star) &- \tfrac{\gamma_k^{-1} - \beta_k}{2}\|x_{k+1} - x_k\|_2^2 + g(x^\star) \\ &+ \tfrac{1}{2\gamma_k}(\|x_k - x^\star\|_2^2 - \|x_{k+1} - x^\star\|_2^2) \end{aligned}$$

which, by multiplying by $2\gamma_k$ and using $p^\star = f(x^\star) + g(x^\star)$, gives

$$\begin{aligned} \|x_{k+1} - x^\star\|_2^2 \leq \|x_k - x^\star\|_2^2 &+ (\beta_k\gamma_k - 1)\|x_{k+1} - x_k\|_2^2 \\ &- 2\gamma_k(f(x_{k+1}) + g(x_{k+1}) - p^\star) \end{aligned}$$

Lyapunov inequality – Convex setting

- The last inequality on previous slide is Lyapunov inequality

$$\begin{aligned}\underbrace{\|x_{k+1} - x^\star\|_2^2}_{V_{k+1}} \leq \underbrace{\|x_k - x^\star\|_2^2}_{V_k} + \underbrace{(\beta_k\gamma_k - 1)\|x_{k+1} - x_k\|_2^2}_{W_k} \\ - 2\gamma_k \underbrace{(f(x_{k+1}) + g(x_{k+1}) - p^\star)}_{R_k}\end{aligned}$$

- Will divide analysis two cases: Short and long step-sizes
  - Step-sizes $\gamma_k \in [\epsilon, \frac{1}{\beta_k}]$: gives $\beta_k\gamma_k \leq 1$ and $W_k \leq 0$
  - Step-sizes $\gamma_k \in [\frac{1}{\beta_k}, \frac{2}{\beta_k} - \epsilon]$: gives $\beta_k\gamma_k \geq 1$ and $W_k \geq 0$

  since W_k contribute differently

Short step-sizes

- For step-sizes $\gamma_k \in [\epsilon, \frac{1}{\beta_k}]$, the Lyapunov inequality implies:

$$\underbrace{\|x_{k+1} - x^\star\|_2^2}_{V_{k+1}} \leq \underbrace{\|x_k - x^\star\|_2^2}_{V_k} - 2\gamma_k \underbrace{(f(x_{k+1}) + g(x_{k+1}) - p^\star)}_{R_k}$$

  where we have used $W_k = 0$ (which is OK since $W_k \leq 0$)
- Nonconvex analysis says function value decreases in every iteration
- Consequences:
  - Distance to solution $\|x_k - x^\star\|_2$ converges as $k \to \infty$
  - Function value decreases to optimal function value as:

$$f(x_{k+1}) + g(x_{k+1}) - p^\star \leq \frac{\|x_0 - x^\star\|_2^2}{2\sum_{i=0}^k \gamma_i}$$

    if f is β-smooth and $\gamma_k = \frac{1}{\beta}$, then converges as $O(1/k)$:

$$f(x_{k+1}) + g(x_{k+1}) - p^\star \leq \frac{\beta\|x_0 - x^\star\|_2^2}{2(k+1)}$$

Long step-sizes

- For step-sizes $\gamma_k \in [\frac{1}{\beta_k}, \frac{2}{\beta_k} - \epsilon]$, the Lyapunov inequality is:

$$\underbrace{\|x_{k+1} - x^\star\|_2^2}_{V_{k+1}} \leq \underbrace{\|x_k - x^\star\|_2^2}_{V_k} + \underbrace{(\beta_k\gamma_k - 1)\|x_{k+1} - x_k\|_2^2}_{W_k}$$
$$- 2\gamma_k \underbrace{(f(x_{k+1}) + g(x_{k+1}) - p^\star)}_{R_k}$$

- From nonconvex analysis can conclude that W_k is summable
  - We showed for $\gamma_k \in [\epsilon, \frac{2}{\beta_k} - \epsilon]$, $(\|x_{k+1} - x_k\|_2^2)_{k\in\mathbb{N}}$ is summable
  - Since $\beta_k\gamma_k$ bounded, also $(W_k)_{k\in\mathbb{N}}$ is summable
  - Let us define $\overline{W} = \sum_{k=0}^{\infty} W_k$
- Consequences:
  - Distance to solution $\|x_k - x^\star\|_2$ converges as $k \to \infty$
  - Function value decreases to optimal function value as:

$$f(x_{k+1}) + g(x_{k+1}) - p^\star \leq \frac{\|x_0 - x^\star\|_2^2 + \overline{W}}{2\sum_{i=0}^{k} \gamma_i}$$

for β-smooth f with $\gamma_k = \frac{1}{\beta}$, denominator replaced by $\frac{2(k+1)}{\beta}$

Outline

- A fundamental inequality
- Nonconvex setting
- Convex setting
- **Strongly convex setting**
- Backtracking
- Stopping conditions
- Accelerated gradient method
- Scaling

Strongly convex setting

- We will analyze the proximal gradient method

$$x_{k+1} = \mathrm{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k))$$

  in a strongly convex setting for solving

$$\mathrm{minimize}\, f(x) + g(x)$$

- Will show linear convergence for distance to solution $\|x_k - x^\star\|_2$
- Two ways to show linear convergence, we can:
  (i) Base analysis on *A fundamental inequality*
  (ii) Start by $\|x_{k+1} - x^\star\|_2^2$ and expand (which is what we will do)

Strongly convex setting – Assumptions

(i)	$f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and σ-strongly convex
(ii)	f is β-smooth
(iii)	$g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is closed convex
(iv)	A minimizer $x^\star$ exists and $p^\star = f(x^\star) + g(x^\star)$ is optimal value
(v)	Algorithm parameters $\gamma_k \in [\epsilon, \frac{2}{\beta} - \epsilon]$, where $\epsilon > 0$

- Assumptions as for fundamental inequality plus
  - σ-strong convexity of f
  - β-smoothness of f instead of upper bound for x_{k+1} and x_k
  - restricted step-size parameters γ_k (as in (non)convex setting)
- But will not use fundamental inequality in analysis

Strongly convex setting – Analysis

Use that

(a) $x^\star = \mathrm{prox}_{\gamma g}(x^\star - \gamma\nabla f(x^\star))$ for all $\gamma > 0$
(b) the proximal operator is nonexpansive
(c) gradients of β-smooth σ-strongly convex functions f satisfy

$$(\nabla f(x) - \nabla f(y))^T(x-y) \geq \tfrac{1}{\beta+\sigma}\|\nabla f(x) - \nabla f(y)\|_2^2 + \tfrac{\sigma\beta}{\beta+\sigma}\|x-y\|_2^2$$

to get

$$\begin{aligned}
&\|x_{k+1} - x^\star\|_2^2 \\
&\quad\overset{(a)}{=} \|\mathrm{prox}_{\gamma_k g}(x_k - \gamma_k\nabla f(x_k)) - \mathrm{prox}_{\gamma_k g}(x^\star - \gamma_k\nabla f(x^\star))\|_2^2 \\
&\quad\overset{(b)}{\leq} \|(x_k - \gamma_k\nabla f(x_k)) - (x^\star - \gamma_k\nabla f(x^\star))\|_2^2 \\
&\quad= \|x_k - x^\star\|_2^2 - 2\gamma_k(\nabla f(x_k) - \nabla f(x^\star))^T(x_k - x^\star) \\
&\qquad + \gamma_k^2\|\nabla f(x_k) - \nabla f(x^\star)\|_2^2 \\
&\quad\overset{(c)}{\leq} \|x_k - x^\star\|_2^2 - \tfrac{2\gamma_k}{\beta+\sigma}(\|\nabla f(x_k) - \nabla f(x^\star)\|_2^2 + \sigma\beta\|x_k - x^\star\|_2^2) \\
&\qquad + \gamma_k^2\|\nabla f(x_k) - \nabla f(x^\star)\|_2^2 \\
&\quad= (1 - \tfrac{2\gamma_k\sigma\beta}{\beta+\sigma})\|x_k - x^\star\|_2^2 - \gamma_k(\tfrac{2}{\beta+\sigma} - \gamma_k)\|\nabla f(x_k) - \nabla f(x^\star)\|_2^2
\end{aligned}$$

Lyapunov inequality – Strongly convex setting

- Lyapunov inequality from previous slide is

$$\begin{aligned}\|x_{k+1} - x^\star\|_2^2 \leq (1 - \tfrac{2\gamma_k\sigma\beta}{\beta+\sigma})\|x_k - x^\star\|_2^2 \\ - \underbrace{\gamma_k(\tfrac{2}{\beta+\sigma} - \gamma_k)\|\nabla f(x_k) - \nabla f(x^\star)\|_2^2}_{W_k}\end{aligned}$$

- Will divide analysis into two cases: Short and long step-sizes
  - Step-sizes $\gamma_k \in [\epsilon, \frac{2}{\beta+\sigma}]$: gives $W_k \geq 0$
  - Step-sizes $\gamma_k \in [\frac{2}{\beta+\sigma}, \frac{2}{\beta} - \epsilon]$: gives $W_k \leq 0$

Short step-sizes

- Lyapunov inequality

$$\begin{aligned}\|x_{k+1}-x^\star\|_2^2 \le (1-\tfrac{2\gamma_k\sigma\beta}{\beta+\sigma})\|x_k-x^\star\|_2^2 \\ -\underbrace{\gamma_k(\tfrac{2}{\beta+\sigma}-\gamma_k)\|\nabla f(x_k)-\nabla f(x^\star)\|_2^2}_{W_k}\end{aligned}$$

  for $\gamma_k \in [\epsilon, \frac{2}{\beta+\sigma}]$ implies $W_k \ge 0$
- Strong monotonicity with modulus σ of ∇f implies

$$\|\nabla f(x_k)-\nabla f(x^\star)\|_2 \ge \sigma\|x_k-x^\star\|_2$$

- So we have linear convergence since

$$\begin{aligned}\|x_{k+1}-x^\star\|_2^2 &\le (1-\tfrac{2\gamma_k\sigma\beta}{\beta+\sigma}-\sigma^2\gamma_k(\tfrac{2}{\beta+\sigma}-\gamma_k))\|x_k-x^\star\|_2^2 \\ &= (1-\tfrac{2\gamma_k\sigma(\beta+\sigma)}{\beta+\sigma}+\sigma^2\gamma_k^2)\|x_k-x^\star\|_2^2 \\ &= (1-\sigma\gamma_k)^2\|x_k-x^\star\|_2^2\end{aligned}$$

  where $(1-\sigma\gamma_k)^2 \in [0,1)$ for full range of γ_k

Long step-sizes

- Lyapunov inequality

$$\|x_{k+1} - x^\star\|_2^2 \le (1 - \tfrac{2\gamma_k \sigma\beta}{\beta+\sigma})\|x_k - x^\star\|_2^2 \\ - \underbrace{\gamma_k(\tfrac{2}{\beta+\sigma} - \gamma_k)\|\nabla f(x_k) - \nabla f(x^\star)\|_2^2}_{W_k}$$

  for $\gamma_k \in [\frac{2}{\beta+\sigma}, \frac{2}{\beta} - \epsilon]$ implies $W_k \le 0$
- That f is β-smooth implies ∇f is β-Lipschitz continuous:

$$\|\nabla f(x_k) - \nabla f(x^\star)\|_2 \le \beta\|x_k - x^\star\|_2$$

- So we have linear convergence since

$$\begin{aligned} \|x_{k+1} - x^\star\|_2^2 &\le (1 - \tfrac{2\gamma_k\sigma\beta}{\beta+\sigma} - \beta^2\gamma_k(\tfrac{2}{\beta+\sigma} - \gamma_k))\|x_k - x^\star\|_2^2 \\ &= (1 - \tfrac{2\gamma_k\beta(\sigma+\beta)}{\beta+\sigma} + \beta^2\gamma_k^2)\|x_k - x^\star\|_2^2 \\ &= (1 - \beta\gamma_k)^2\|x_k - x^\star\|_2^2 \end{aligned}$$

  where $(1 - \beta\gamma_k)^2 \in [0, 1)$ for full range of γ_k

Unified rate

- By removing the square and checking sign, we have
  - for step-sizes $\gamma_k \in [\epsilon, \frac{2}{\beta+\sigma}]$:

$$\|x_{k+1} - x^\star\|_2 \le (1 - \sigma\gamma_k)\|x_k - x^\star\|_2$$

  - for step-sizes $\gamma_k \in [\frac{2}{\beta+\sigma}, \frac{2}{\beta} - \epsilon]$:

$$\|x_{k+1} - x^\star\|_2 \le (\beta\gamma_k - 1)\|x_k - x^\star\|_2$$

- The linear convergence result can be summarized as

$$\|x_{k+1} - x^\star\|_2 \le \max(1 - \sigma\gamma_k, \beta\gamma_k - 1)\|x_k - x^\star\|_2$$

Optimal step-size

- For fixed-step-sizes $\gamma_k = \gamma$, the rate result is

$$\|x_{k+1} - x^\star\|_2 \leq \underbrace{\max(1 - \sigma\gamma, \beta\gamma - 1)}_{\rho} \|x_k - x^\star\|_2$$

- Optimal γ that gives smallest contraction is $\gamma = \frac{2}{\beta+\sigma}$:
  - $(1 - \sigma\gamma)$ decreasing in γ, optimal at upper bound $\gamma = \frac{2}{\beta+\sigma}$
  - $(\beta\gamma - 1)$ increasing in γ, optimal at lower bound $\gamma = \frac{2}{\beta+\sigma}$
  - Bounds coincide at $\gamma = \frac{2}{\beta+\sigma}$ to give rate factor $\rho = \frac{\beta-\sigma}{\beta+\sigma}$

Outline

- A fundamental inequality
- Nonconvex setting
- Convex setting
- Strongly convex setting
- **Backtracking**
- Stopping conditions
- Accelerated gradient method
- Scaling

Choose β_k and γ_k

- In nonconvex and convex analysis, we assume β_k known such that

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \tfrac{\beta_k}{2}\|x_k - x_{k+1}\|_2^2$$

  for consecutive iterates x_k and x_{k+1}
- This is an assumption on the function f
- We call it *descent condition* (DC)
- If f is β-smooth, then $\beta_k = \beta$ is valid choice since

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \tfrac{\beta}{2}\|x - y\|_2^2$$

  for all x, y, then we can select $\gamma_k \in [\epsilon, \frac{2}{\beta} - \epsilon]$

Choose β_k and γ_k – Backtracking

- Backtracking: choose $\kappa > 1$, $\beta_{k,0} \in [\eta, \eta^{-1}]$, let $l_k = 0$, and loop
  1. choose $\gamma_k \in [\epsilon, \frac{2}{\beta_{k,l_k}} - \epsilon]$
  2. compute $x_{k+1} = \text{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k))$
  3. **if** descent condition (DC) satisfied

     set $k \leftarrow k+1$ `// increment algorithm counter`

     set $\bar{l}_k \leftarrow l_k$ `// store final backtrack counter`

     set $\beta_k \leftarrow \beta_{k,l_k}$ `// store final` β `variable`

     break backtrack loop

     else

     set $\beta_{k,l_k+1} \leftarrow \kappa\beta_{k,l_k}$ `// increase backtrack parameter`

     set $l_k \leftarrow l_k + 1$ `// increment backtrack counter`

     end
- Larger β_{k,l_k} gives smaller upper bound for step-size γ_k
- Forwardtracking on β_{k,l_k}, backtracking for γ_k upper bound

When to use backtracking

- f is β-smooth but constant β unknown:
  - initialize $\beta_{k,0} = \beta_{k-1,\bar{l}_{k-1}}$ to previously used value
  - then $(\beta_k)_{k\in\mathbb{N}}$ nondecreasing
  - finally $\beta_k \geq \beta$ (if needed), then
    - step-size bound $\gamma_k \in [\epsilon, \frac{2}{\beta_{k,\bar{l}_k}} - \epsilon]$ makes (DC) hold directly
    - so will have constant β_k after finite number of algoritm iterations
- ∇f locally Lipschitz and sequence bounded (as in convex case):
  - initialize $\beta_{k,0} = \bar{\beta}$, for some pre-chosen $\bar{\beta} > 0$
  - reset to same value $\bar{\beta}$ in every algorithm iteration
  - will find a local Lipschitz constant

Outline

- A fundamental inequality
- Nonconvex setting
- Convex setting
- Strongly convex setting
- Backtracking
- **Stopping conditions**
- Accelerated gradient method
- Scaling

When to stop algorithm?

- Consider $\underset{x}{\mathrm{minimize}}\, f(x) + g(x)$
- Apply proximal gradient method $x_{k+1} = \mathrm{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k))$
- Algorithm sequence satisfies

$$\partial g(x_{k+1}) + \nabla f(x_{k+1}) \ni \underbrace{\gamma_k^{-1}(x_k - x_{k+1}) + \nabla f(x_{k+1}) - \nabla f(x_k)}_{u_k} \to 0$$

is $\|u_k\|_2$ small a good measure of being close to fixed-point?

When to stop algorithm – Scaled problem

Let $a > 0$ and solve equivalent problem $\underset{x}{\text{minimize}}\, af(x) + ag(x)$:

- Denote algorithm parameter $\gamma_{a,k} = \frac{\gamma_k}{a}$
- Algorithm satisfies:

$$x_{k+1} = \text{prox}_{\gamma_{a,k} ag}(x_k - \gamma_{a,k} \nabla af(x_k)) = \text{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k))$$

  i.e., the same algorithm as before
- However, $u_{a,k}$ in this setting satisfies

$$\begin{aligned} u_{a,k} &= \gamma_{a,k}^{-1}(x_k - x_{k+1}) + \nabla af(x_{k+1}) - \nabla af(x_k) \\ &= a(\gamma_k^{-1}(x_k - x_{k+1}) + \nabla f(x_{k+1}) - \nabla f(x_k)) \\ &= au_k \end{aligned}$$

  i.e., same algorithm but different optimality measure
- Optimality measure should be scaling invariant

Scaling invariant stopping condition

- For β-smooth f, use scaled condition $\frac{1}{\beta}u_k$

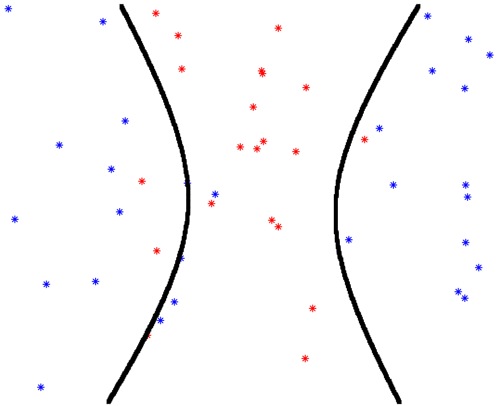
$$\frac{1}{\beta}u_k := \frac{1}{\beta}(\gamma_k^{-1}(x_k - x_{k+1}) + \nabla f(x_{k+1}) - \nabla f(x_k))$$

  that we have seen before
- Let us scale problem by a to get $\mathrm{minimize}\, af(x) + ag(x)$, then
  - smoothness constant $\beta_a = a\beta$ scaled by $a \Rightarrow$ use $\gamma_{a,k} = \frac{\gamma_k}{a}$
  - optimality measure $\frac{1}{\beta_a}u_{a,k} = \frac{1}{a\beta}au_k = \frac{1}{\beta}u_k$ remains the same

  so it is scaling invariant
- Problem considered solved to optimality if, say, $\frac{1}{\beta}\|u_k\|_2 \leq 10^{-6}$
- Often lower accuracy 10^{-3} to 10^{-4} is enough

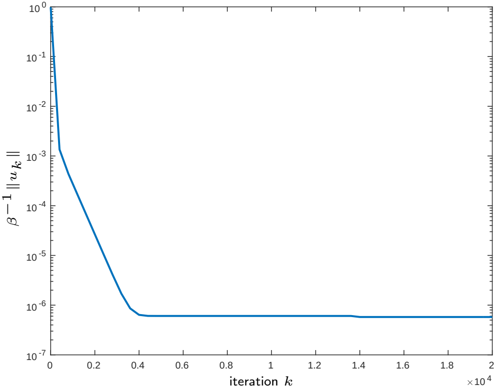
Example – SVM

- Classification problem from SVM lecture, SVM with
  - polynomial features of degree 2
  - regularization parameter $\lambda = 0.00001$



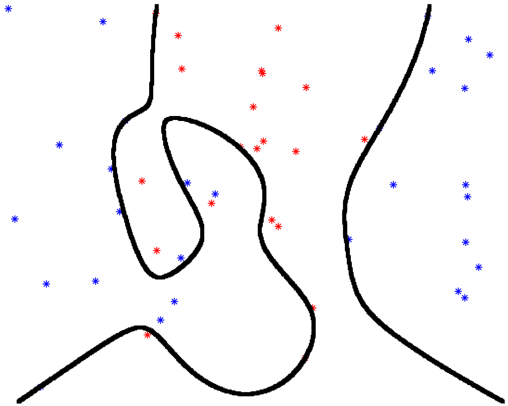
Example – Optimality measure

- Plots $\beta^{-1}\|u_k\|_2 = \beta^{-1}\|\gamma_k^{-1}(x_k - x_{k+1}) + \nabla f(x_{k+1}) - \nabla f(x_k)\|_2$
- Shows $\beta^{-1}\|u_k\|_2$ up to 20'000 iterations
- Quite many iterations needed to converge



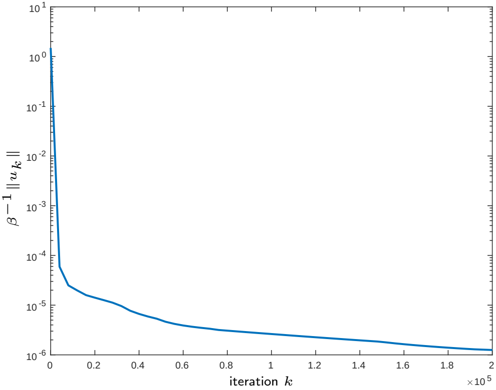
Example – SVM higher degree polynomial

- Classification problem from SVM lecture, SVM with
  - polynomial features of degree 6
  - regularization parameter $\lambda = 0.00001$



Example – Optimality measure

- Plots $\beta^{-1}\|u_k\|_2 = \beta^{-1}\|\gamma_k^{-1}(x_k - x_{k+1}) + \nabla f(x_{k+1}) - \nabla f(x_k)\|_2$
- Shows $\beta^{-1}\|u_k\|_2$ up to 200'000 iterations (10x more than before)
- Many iterations needed for high accuracy



Outline

- A fundamental inequality
- Nonconvex setting
- Convex setting
- Strongly convex setting
- Backtracking
- Stopping conditions
- **Accelerated gradient method**
- Scaling

Accelerated proximal gradient method

- Consider *convex* composite problem

$$\underset{x}{\text{minimize}}\, f(x) + g(x)$$

  where
  - $f : \mathbb{R}^n \to \mathbb{R}$ is β-smooth and convex
  - $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is closed and convex
- Proximal gradient descent

$$x_{k+1} = \text{prox}_{\gamma g}(x_k - \gamma \nabla f(x_k))$$

  achieves $O(1/k)$ convergence rate in function value
- *Accelerated* proximal gradient method

$$\begin{aligned} y_k &= x_k + \theta_k(x_k - x_{k-1}) \\ x_{k+1} &= \text{prox}_{\gamma g}(y_k - \gamma \nabla f(y_k)) \end{aligned}$$

  (with specific θ_k) achieves faster $O(1/k^2)$ convergence rate

Accelerated proximal gradient method – Parameters

- *Accelerated* proximal gradient method

$$y_k = x_k + \theta_k(x_k - x_{k-1})$$
$$x_{k+1} = \mathrm{prox}_{\gamma g}(y_k - \gamma\nabla f(y_k))$$

- Step-sizes are restricted $\gamma \in (0, \frac{1}{\beta}]$
- The θ_k parameters can be chosen either as

$$\theta_k = \tfrac{k-1}{k+2}$$

  or $\theta_k = \frac{t_{k-1}-1}{t_k}$ where

$$t_k = \tfrac{1+\sqrt{1+4t_{k-1}^2}}{2}$$

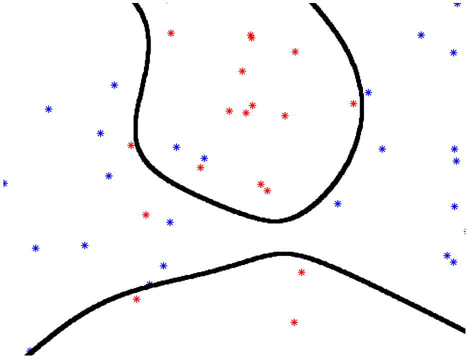
  these choices are very similar
- Algorithm behavior in nonconvex setting not well understood

Not a descent method

- Descent method means function value is decreasing every iteration
- We know that proximal gradient method is a descent method
- However, accelerated proximal gradient method is not

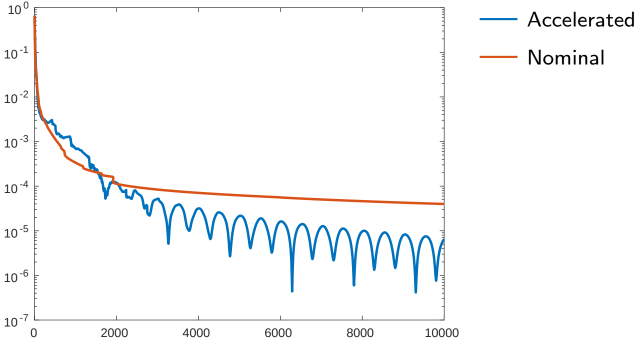
Accelerated gradient method – Example

- Accelerated vs nominal proximal gradient method
- Problem from SVM lecture, polynomial deg 6 and $\lambda = 0.0215$



Accelerated gradient method – Example

- Accelerated vs nominal proximal gradient method
- Problem from SVM lecture, polynomial deg 6 and $\lambda = 0.0215$



Outline

- A fundamental inequality
- Nonconvex setting
- Convex setting
- Strongly convex setting
- Backtracking
- Stopping conditions
- Accelerated gradient method
- **Scaling**

Scaled proximal gradient method

- Proximal gradient method:

$$x_{k+1} = \operatorname*{argmin}_y \left(\underbrace{f(x_k) + \nabla f(x_k)^T(y - x) + \tfrac{1}{2\gamma_k}\|y - x_k\|_2^2}_{\hat{f}_{x_k}(y)} + g(y) \right)$$

  approximates function $f(y)$ around x_k by $\hat{f}_{x_k}(y)$
- The better the approximation, the faster the convergence
- By scaling: we mean to use an approximation of the form

$$\hat{f}_{x_k}(y) = f(x_k) + \nabla f(x_k)^T(y - x_k) + \tfrac{1}{2\gamma_k}\|y - x_k\|_H^2$$

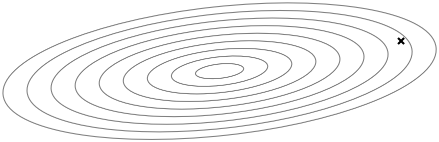
  where $H \in \mathbb{R}^{n\times n}$ is a positive definite matrix and $\|x\|_H^2 = x^T H x$

Gradient descent – Example

- Gradient descent on β-smooth quadratic problem

$$\underset{x}{\text{minimize}} \;\; \frac{1}{2}\begin{bmatrix}x_1\\x_2\end{bmatrix}^T\begin{bmatrix}0.1 & -0.1\\-0.1 & 1\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}$$

- Step-size $\gamma = \frac{1}{\beta}$ and norm $\|\cdot\|_2$ in model

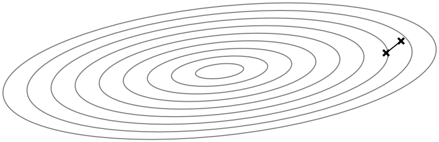


Gradient descent – Example

- Gradient descent on β-smooth quadratic problem

$$\underset{x}{\text{minimize}} \ \frac{1}{2}\begin{bmatrix}x_1\\x_2\end{bmatrix}^T\begin{bmatrix}0.1 & -0.1\\-0.1 & 1\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}$$

- Step-size $\gamma = \frac{1}{\beta}$ and norm $\|\cdot\|_2$ in model

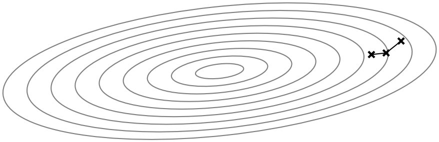


Gradient descent – Example

- Gradient descent on β-smooth quadratic problem

$$\underset{x}{\text{minimize}}\ \frac{1}{2}\begin{bmatrix}x_1\\x_2\end{bmatrix}^T\begin{bmatrix}0.1 & -0.1\\-0.1 & 1\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}$$

- Step-size $\gamma = \frac{1}{\beta}$ and norm $\|\cdot\|_2$ in model

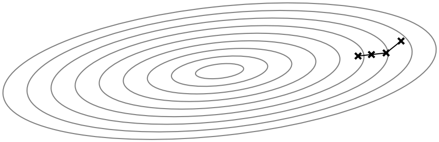


Gradient descent – Example

- Gradient descent on β-smooth quadratic problem

$$\underset{x}{\text{minimize}} \;\; \frac{1}{2}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size $\gamma = \frac{1}{\beta}$ and norm $\|\cdot\|_2$ in model

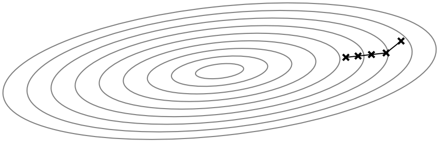


Gradient descent – Example

- Gradient descent on β-smooth quadratic problem

$$\underset{x}{\text{minimize}} \; \frac{1}{2}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size $\gamma = \frac{1}{\beta}$ and norm $\|\cdot\|_2$ in model

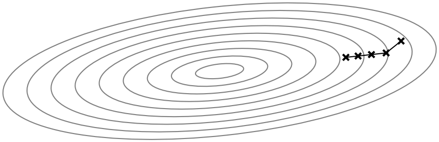


Gradient descent – Example

- Gradient descent on β-smooth quadratic problem

$$\underset{x}{\text{minimize}} \;\; \frac{1}{2}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size $\gamma = \frac{1}{\beta}$ and norm $\|\cdot\|_2$ in model

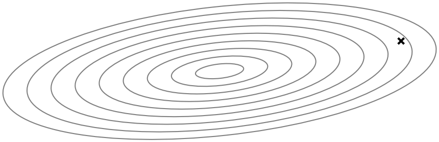


Scaled gradient descent – Example

- Gradient descent on β-smooth quadratic problem

$$\underset{x}{\text{minimize}} \;\; \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Scaling $H = \mathbf{diag}(\nabla^2 f)$, γ is inverse smoothness w.r.t. $\|\cdot\|_H$

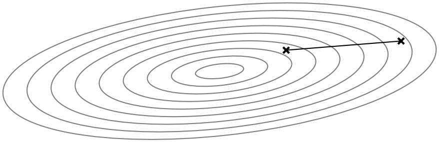


Scaled gradient descent – Example

- Gradient descent on β-smooth quadratic problem

$$\underset{x}{\text{minimize}} \;\; \frac{1}{2}\begin{bmatrix}x_1\\x_2\end{bmatrix}^T \begin{bmatrix}0.1 & -0.1\\-0.1 & 1\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}$$

- Scaling $H = \mathbf{diag}(\nabla^2 f)$, γ is inverse smoothness w.r.t. $\|\cdot\|_H$

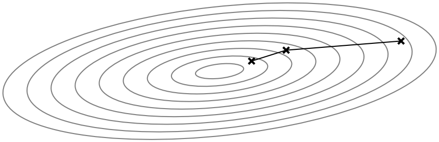


Scaled gradient descent – Example

- Gradient descent on β-smooth quadratic problem

$$\underset{x}{\text{minimize}} \ \frac{1}{2}\begin{bmatrix}x_1\\x_2\end{bmatrix}^T\begin{bmatrix}0.1 & -0.1\\-0.1 & 1\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}$$

- Scaling $H = \mathbf{diag}(\nabla^2 f)$, γ is inverse smoothness w.r.t. $\|\cdot\|_H$

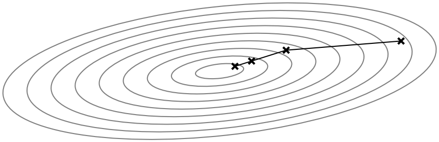


Scaled gradient descent – Example

- Gradient descent on β-smooth quadratic problem

$$\underset{x}{\text{minimize}} \;\; \frac{1}{2}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Scaling $H = \mathbf{diag}(\nabla^2 f)$, γ is inverse smoothness w.r.t. $\|\cdot\|_H$

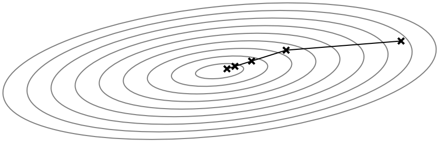


Scaled gradient descent – Example

- Gradient descent on β-smooth quadratic problem

$$\underset{x}{\text{minimize}} \;\; \frac{1}{2}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Scaling $H = \mathbf{diag}(\nabla^2 f)$, γ is inverse smoothness w.r.t. $\|\cdot\|_H$

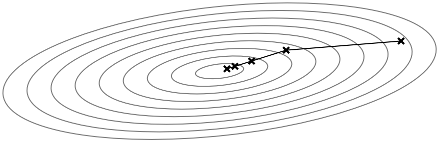


Scaled gradient descent – Example

- Gradient descent on β-smooth quadratic problem

$$\underset{x}{\text{minimize}} \;\; \frac{1}{2}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Scaling $H = \mathbf{diag}(\nabla^2 f)$, γ is inverse smoothness w.r.t. $\|\cdot\|_H$



Smoothness w.r.t. $\|\cdot\|_H$

What is $\|\cdot\|_H$?

- Requirement: $H \in \mathbb{R}^{n\times n}$ is symmetric positive definite ($H \succ 0$)
- The norm $\|x\|_H^2 := x^T H x$, for $H = I$, we get $\|x\|_I^2 = \|x\|_2^2$

Smoothness

- Function $f : \mathbb{R}^n \to \mathbb{R}$ is β-smooth if for all $x, y \in \mathbb{R}^n$:

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \tfrac{\beta}{2}\|x - y\|_2^2$$
$$f(y) \geq f(x) + \nabla f(x)^T (y - x) - \tfrac{\beta}{2}\|x - y\|_2^2$$

- We say f β_H-smoothness w.r.t. scaled norm $\|\cdot\|_H$ if

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \tfrac{\beta_H}{2}\|x - y\|_H^2$$
$$f(y) \geq f(x) + \nabla f(x)^T (y - x) - \tfrac{\beta_H}{2}\|x - y\|_H^2$$

  for all $x, y \in \mathbb{R}^n$
- If f is smooth (w.r.t. $\|\cdot\|_2$) it is also smooth w.r.t. $\|\cdot\|_H$

Example – A quadratic

- Let $f(x) = \frac{1}{2}x^T Hx = \frac{1}{2}\|x\|_H^2$ with $H \succ 0$
- f is 1-smooth w.r.t $\|\cdot\|_H$ (with equality):

$$\begin{aligned} f(x) &+ \nabla f(x)^T(y-x) + \tfrac{1}{2}\|x-y\|_H^2 \\ &= \tfrac{1}{2}x^T Hx + (Hx)^T(y-x) + \tfrac{1}{2}\|x-y\|_H^2 \\ &= \tfrac{1}{2}x^T Hx + (Hx)^T(y-x) + \tfrac{1}{2}(\|x\|_H^2 - 2(Hx)^T y + \|y\|_H^2) \\ &= \tfrac{1}{2}\|y\|_H^2 = f(y) \end{aligned}$$

  which holds also if adding linear term $q^T x$ to f

- f is $\lambda_{\max}(H)$-smooth (w.r.t. $\|\cdot\|_2$), continue equality:

$$\begin{aligned} f(y) &= f(x) + \nabla f(x)^T(y-x) + \tfrac{1}{2}\|x-y\|_H^2 \\ &\leq f(x) + \nabla f(x)^T(y-x) + \tfrac{\lambda_{\max}(H)}{2}\|x-y\|_2^2 \end{aligned}$$

  much more conservative estimate of function!

Scaled proximal gradient for quadratics

- Let $f(x) = \frac{1}{2}x^T Hx$ with $H \succ 0$, which is 1-smooth w.r.t. $\|\cdot\|_H$
- Approximation with scaled norm $\|\cdot\|_H$ and $\gamma_k = 1$ satisfies $\forall x_k$:

$$\hat{f}_{x_k}(y) = f(x_k) + \nabla f(x_k)^T(y - x_k) + \tfrac{1}{2}\|x_k - y\|_H^2 = f(y)$$

  since f is 1-smooth w.r.t. $\|\cdot\|_H$ with equality
- An iteration then reduces to solving problem itself:

$$x_{k+1} = \underset{y}{\operatorname{argmin}}(\hat{f}_{x_k}(y) + g(y)) = \underset{y}{\operatorname{argmin}}(f(y) + g(y))$$

- Model very accurate, but very expensive iterations

Scaled proximal gradient method reformulation

- Proximal gradient method with scaled norm $\|\cdot\|_H$:

$$\begin{aligned} x_{k+1} &= \operatorname*{argmin}_y \Big(f(x_k) + \nabla f(x_k)^T(y-x) + \tfrac{1}{2\gamma_k}\|y-x_k\|_H^2 + g(y)\Big) \\ &= \operatorname*{argmin}_y \Big(g(y) + \tfrac{1}{2\gamma_k}\|y-(x_k-\gamma_k H^{-1}\nabla f(x_k)\|_H^2\Big) \\ &=: \operatorname{prox}_{\gamma_k g}^H(x_k-\gamma_k H^{-1}\nabla f(x_k)) \end{aligned}$$

  where $H = I$ gives nominal method

- Computational difference per iteration:
  1. Need to invert H^{-1} (or solve $Hd_k = \nabla f(x_k)$)
  2. Need to compute prox with new metric

$$\operatorname{prox}_{\gamma_k g}^H(z) := \operatorname*{argmin}_x(g(x) + \tfrac{1}{2\gamma_k}\|x-z\|_H^2)$$

     that may be very costly

Computational cost

- Assume that H is dense or general sparse
  - H^{-1} dense: cubic complexity (vs maybe quadratic for gradient)
  - H^{-1} sparse: lower than cubic complexity
  - $\mathrm{prox}_{\gamma_k g}^{H}$: difficult optimization problem
- Assume that H is diagonal
  - H^{-1}: invert diagonal elements – linear complexity
  - $\mathrm{prox}_{\gamma_k g}^{H}$: often as cheap as nominal prox (e.g., for separable g)
  - this gives individual step-sizes for each coordinate
- Assume that H is block-diagonal with small blocks
  - H^{-1}: invert individual blocks – also cheap
  - $\mathrm{prox}_{\gamma_k g}^{H}$: often quite cheap (e.g., for block-separable g)
- If $H = I$, method is nominal method

Convergence

- We get similar results as in the nominal $H = I$ case
- We assume β_H smoothness w.r.t. $\|\cdot\|_H$
- We can replace all $\|\cdot\|_2$ with $\|\cdot\|_H$ and ∇f with $H^{-1}\nabla f$:
  - Nonconvex setting with $\gamma_k = \frac{1}{\beta_H}$

$$\min_{l \in \{0,\dots,k\}} \|\nabla f(x_l)\|_{H^{-1}}^2 \leq \frac{2\beta_H(f(x_0) + g(x_0) - p^\star)}{k+1}$$

  - Convex setting with $\gamma_k = \frac{1}{\beta_H}$

$$f(x_k) + g(x_k) - p^\star \leq \frac{\beta_H \|x_0 - x^\star\|_H^2}{2(k+1)}$$

  - Strongly convex setting with f σ_H-strongly convex w.r.t. $\|\cdot\|_H$

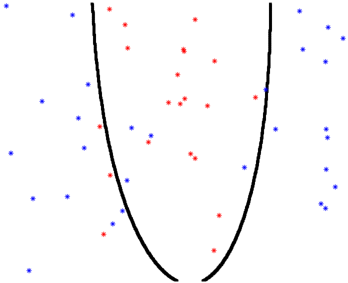
$$\|x_{k+1} - x^\star\|_H \leq \max(\beta_H\gamma - 1, 1 - \sigma_H\gamma)\|x_k - x^\star\|_H$$

Example – Logistic regression

- Logistic regression with $\theta = (w, b)$:

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{N} \log(1 + e^{w^T\phi(x_i)+b}) - y_i(w^T\phi(x_i) + b) + \tfrac{\lambda}{2}\|w\|_2^2$$

  on the following data set (from logistic regression lecture)
- Polynomial features of degree 6, Tikhonov regularization $\lambda = 0.01$
- Number of decision variables: 28



Algorithms

Compare the following algorithms, all with backtracking:

1. Gradient method
2. Gradient method with fixed diagonal scaling
3. Gradient method with fixed full scaling

Fixed scalings

- Logistic regression gradient and Hessian satisfy with $L = [X, \mathbf{1}]$

$$\nabla f(\theta) = L^T(\sigma(L\theta) - Y) + \lambda I_w \theta \quad \nabla^2 f(\theta) = L^T \sigma'(L\theta) L + \lambda I_w$$

  where σ is the (vector-version of) sigmoid, and $I_w(w, b) = (w, 0)$
- The sigmoid function σ is 0.25-Lipschitz continuous
- Gradient method with fixed full scaling (3.) uses

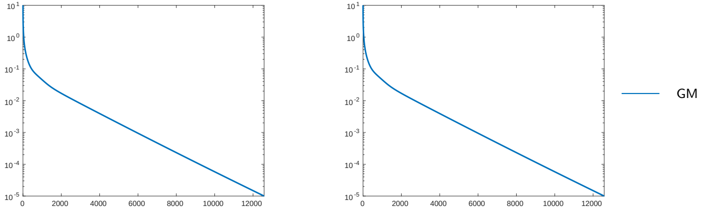
$$H = 0.25 L^T L + \lambda I_w$$

- Gradient method with fixed diagonal scaling (2.) uses

$$H = \mathbf{diag}(0.25 L^T L + \lambda I_w)$$

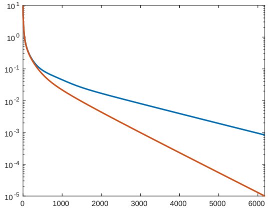
Example – Numerics

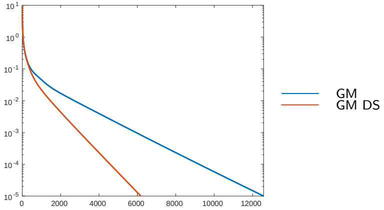
- Logistic regression polynomial features of degree 6, $\lambda = 0.01$
- Standard gradient method with backtracking (GM)



Example – Numerics

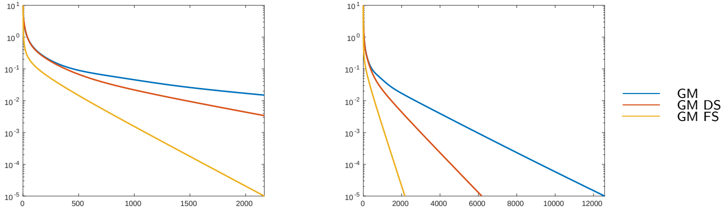
- Logistic regression polynomial features of degree 6, $\lambda = 0.01$
- Gradient method with diagonal scaling (GM DS)





Example – Numerics

- Logistic regression polynomial features of degree 6, $\lambda = 0.01$
- Gradient method with full matrix scaling (GM FS)



Comments

- Smaller number of iterations with better scaling
- Performance is roughly (iteration cost)$\times$(number of iterations)
  - We have only compared number of iterations
  - Iteration cost for (GM) and (GM DS) are the same
  - Iteration cost for (GM FS) higher
  - Need to quantify iteration cost to assess which is best
- In general, can be difficult to find H that performs better