1. Introduction

Miscellaneous

- $\forall x \in \mathbb{R}, 1 + x \le e^x$
- Cosine Thm: $2\mathbf{v}^\top \mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v}-\mathbf{w}\|^2$
- For a random vector $\mathbf{X}$, $\mathrm{Var}(\mathbf{X}) := \mathbb{E}\left[\|\mathbf{X}-\mathbb{E}(\mathbf{X})\|^2\right] = \mathbb{E}\left[\|\mathbf{X}\|^2\right] - \|\mathbb{E}[\mathbf{X}]\|^2 \Rightarrow \mathbb{E}\left[\|\mathbf{X}-\mathbb{E}(\mathbf{X})\|^2\right] \le \mathbb{E}\left[\|\mathbf{X}\|^2\right]$
- $\sum_{t=1}^T 1/\sqrt{t} = \mathcal{O}(\sqrt{T})$
- $\sum_{t=1}^T 1/t = \mathcal{O}(\ln T)$

Eigendecomposition

- Square $n \times n$ matrix $\mathbf{A}$. $\mathbf{A}\mathbf{v} = \lambda \mathbf{v}$
- Equation for eigenvalues: $p(\lambda) = \det(\mathbf{A} - \lambda \mathbf{I}) = 0$
- $\mathbf{A}$ can be factorized as $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$
- If none of eigenvalues are zero, then $\mathbf{A}$ is invertible and its inverse is given by $\mathbf{A}^{-1} = \mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^{-1}$
- $\mathbf{A}^n = \mathbf{Q}\mathbf{\Lambda}^n\mathbf{Q}^{-1}$, $\mathbf{A}^{-1} = \mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^{-1}$

Eigendecomposition of Symmetric Matrices

- For every $n \times n$ real symmetric matrix, $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$, $\mathbf{V}\mathbf{V}^\top = \mathbf{V}^\top\mathbf{V} = \mathbf{I}_n$, $\mathbf{V}^\top = \mathbf{V}^{-1}$
- Can be written as $\mathbf{A} = \sum_{i=1}^n \lambda_i \mathbf{v}_i, \mathbf{v}_i^\top$, $\{\mathbf{v}_i\}_{i=1}^n$ is a basis of $\mathbb{R}^n$.
- Any vector $x \in \mathbb{R}^n$ can be written as $\mathbf{x} = \sum_{i=1}^n \alpha_i \mathbf{v}_i$ for a unique set of $\{\alpha_i\}_{i=1}^n$. Then $\mathbf{A}\mathbf{x} = \sum_{i=1}^n \lambda_i \alpha_i v_i$, $\mathbf{x}^\top \mathbf{A}\mathbf{x} = \sum_{i=1}^n \alpha_i^2 \lambda_i$
- $\max\limits_{\|\mathbf{x}\|=1} \mathbf{x}^\top \mathbf{A}\mathbf{x} = \max\limits_{i=1,\ldots,n} \{\lambda_i\}$, $\min\limits_{\|\mathbf{x}\|=1} \mathbf{x}^\top \mathbf{A}\mathbf{x} = \min\limits_{i=1,\ldots,n} \{\lambda_i\}$
- $\mathbf{A}^k = \sum_{i=1}^n \lambda_i^k \mathbf{v}_i \mathbf{v}_i^\top$, $\mathbf{A}^{-1} = \sum_{i=1}^n \lambda_i^{-1} \mathbf{v}_i \mathbf{v}_i^\top$

Matrix Norm

Spectral Norm

The spectral norm of a matrix $\mathbf{A}$ is the largest singular value of $\mathbf{A}$ (i.e., the square root of the largest eigenvalue of the matrix $\mathbf{A}^*\mathbf{A}$, where $\mathbf{A}^*$ denotes the conjugate transpose $\mathbf{A}$):

$$\|\mathbf{A}\|_2 = \sqrt{\lambda_{\max}(\mathbf{A}^*\mathbf{A})} = \sigma_{\max}(\mathbf{A})$$

For the square matrix,

$$\|\mathbf{A}\|_2 := \max_{\mathbf{v}\in\mathbb{R}^d, \mathbf{v}\neq 0} \frac{\|\mathbf{A}\mathbf{v}\|}{\|\mathbf{v}\|} = \max_{\|\mathbf{v}\|=1} \|\mathbf{A}\mathbf{v}\| = \lambda_{\max}(\mathbf{A})$$

Frobenius Norm

$$\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\sum_{i,j=1}^n |a_{ij}|^2} = \sqrt{\mathrm{tr}(\mathbf{A}^\top\mathbf{A})} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2(\mathbf{A})}$$

Inequalities

- $\|\mathbf{A}\|_2 \le \|\mathbf{A}\|_F$ for every matrix (Proof see Hw8 Ex3).
- $\|\mathbf{A}\,\mathbf{y}\|_2 \le \|\mathbf{A}\|_2 \cdot \|\mathbf{y}\|_2$ for any $\mathbf{A}, \mathbf{y}$
- $\|\mathbf{A}\,\mathbf{x}\|_2^2 \le \lambda_{\max}(\mathbf{A}) \cdot \mathbf{x}^\top \mathbf{A}\,\mathbf{x} = \|\mathbf{A}\|_2 \cdot \mathbf{x}^\top \mathbf{A}\,\mathbf{x}$
- $\|\mathbf{A}\,\mathbf{x}\|_2^2 \ge \lambda_{\min}^2(\mathbf{A}) \cdot \|\mathbf{x}\|_2^2$
- $\lambda_{\min}(\mathbf{A}) \cdot \|\mathbf{x}\|_2^2 \le \mathbf{x}^\top \mathbf{A}\,\mathbf{x} \le \lambda_{\max}(\mathbf{A}) \cdot \|\mathbf{x}\|_2^2 = \|\mathbf{A}\|\|\mathbf{x}\|_2^2$
- If $\lambda_{\min}(\mathbf{A}) \ge \mu$ and/or $\lambda_{\max}(\mathbf{A}) \le L$, then $\lambda_{\max}(\mathbf{A}^{-1}) \le \frac{1}{\mu}$ and/or $\lambda_{\min}(\mathbf{A}^{-1}) \le \frac{1}{L}$
- $\sigma_{i+j-1}(\mathbf{A}+\mathbf{B}) \leqslant \sigma_i(\mathbf{A}) + \sigma_j(\mathbf{B})$ for $i, j \in \mathbb{N}, i+j-1 \le \min\{m,n\}$
- $\|\mathbf{A}\,\mathbf{B}\|_F \ge \sigma_{\min}(\mathbf{B}) \cdot \|\mathbf{A}\|_F$

General Norms and Dual Norms

Norm

A function $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}_+$ is a norm if (a) $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = 0$;
(b) $\|\alpha\mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|$;
(c) $\|\mathbf{x}+\mathbf{y}\| \le \|\mathbf{x}\| + \|\mathbf{y}\|$.

Dual Norm

$$\|\mathbf{y}\|_* := \max_{\|\mathbf{x}\|\le 1} \langle \mathbf{x}, \mathbf{y} \rangle$$

For $p \ge 1$ and $1/p + 1/q = 1$,

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}, \|\cdot\|_{p,*} = \|\cdot\|_q$$

Inequality: $\frac{1}{\sqrt{d}}\|\mathbf{x}\|_2 \le \|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_2 \le \|\mathbf{x}\|_1 \le \sqrt{d}\|\mathbf{x}\|_2$

General Smoothness and Strong Convexity

- Convexity:

$$f(\mathbf{y}) \ge f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$

- Optimality Condition for Convex Functions: Suppose that $f : \mathrm{dom}(f) \to \mathbb{R}$ is convex and differentiable over an open domain $\mathrm{dom}(f) \subseteq \mathbb{R}^d$, and let $X \subseteq \mathrm{dom}(f)$ be a convex set. Point $\mathbf{x}^\star \in X$ is a minimizer of f over X iff

$$\nabla f\left(\mathbf{x}^\star\right)^\top \left(\mathbf{x} - \mathbf{x}^\star\right) \ge 0 \quad \forall \mathbf{x} \in X$$

- Lipschitz continuity:
  $f(\mathbf{x})$ is B-Lipschitz continuous on X if

$$\begin{aligned} & \|f(\mathbf{x}) - f(\mathbf{y})\| \le B\|\mathbf{x}-\mathbf{y}\|_\bullet, \forall \mathbf{x}, \mathbf{y} \in X \\ \Longleftrightarrow\ & \|\mathbf{g}\|_* \le B \text{ for all } \mathbf{g} \in \partial f(\mathbf{x}) \end{aligned}$$

  In particular, $\|f(\mathbf{x}) - f(\mathbf{y})\|_2 \le B\|\mathbf{x}-\mathbf{y}\|_2 \Longleftrightarrow \|\mathbf{g}\|_2 \le B$

- Smoothness:
  - $f(\mathbf{x})$ is L-smooth on X if $f(\mathbf{x})$ is differentiable and
  $$f(\mathbf{x}) \le f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x}-\mathbf{y}) + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2, \forall \mathbf{x}, \mathbf{y} \in X$$
  - $f(\mathbf{x})$ is L-smooth iff $\nabla f(\mathbf{x})$ is L-Lipschitz, i.e.,:
  $$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \le L\|\mathbf{x}-\mathbf{y}\|, \forall \mathbf{x}, \mathbf{y}$$
  - If f is L-smooth then
  $$f\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right) - f(\mathbf{x}) \le -\frac{1}{2L}\|\nabla f(\mathbf{x})\|_2^2,$$
  and
  $$f(\mathbf{x}) - f(\mathbf{x}^*) \ge \frac{1}{2L}\|\nabla f(\mathbf{x})\|_2^2,$$
  (Proof see Hw6 Ex1)
  - In all inequalities above, $\mathbf{y}$ is often set as $\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})$.
- Strong convexity:
  - $f(\mathbf{x})$ is μ-strongly convex on X if
  $$f(\mathbf{x}) \ge f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x}-\mathbf{y}) + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2, \forall \mathbf{x}, \mathbf{y} \in X$$
  - If f is μ–strongly convex, then it also satisfies the PL inequality,
  $$\|\nabla f(\mathbf{x})\|_2^2 \ge 2\mu[f(\mathbf{x}) - f(\mathbf{x}^*)]$$
- Smooth and Convex: If $f(\mathbf{x})$ is convex and L-smooth then
$$f(\mathbf{y}) - f(\mathbf{x}) \le \nabla f(\mathbf{y})^\top (\mathbf{y}-\mathbf{x}) - \frac{1}{2L}\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_2^2$$
$$[\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})]^\top (\mathbf{y}-\mathbf{x}) \ge \frac{1}{L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|$$

Rate of Convergence

Consider the sequence $\{\delta_t\}_t \ge 0$ such that $\delta_0 \ge 0$ and

$$\delta_t - \delta_{t+1} \ge C \cdot \delta_t^\alpha, \quad \forall t \ge 0$$

for some $C > 0$ and $\alpha > 0$.
For any $\alpha > 1$, we have

$$\delta_t = \mathcal{O}\left(\frac{1}{t^{1/(\alpha-1)}}\right)$$

For optimization problems, if we have

$$f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) \ge C \cdot (f(\mathbf{x}_t) - f^*)^\alpha, \quad \forall t \ge 0$$

Then it implies that

- If $\alpha = 1$ and $0 < C < 1$, then $\{\mathbf{x}_t\}_t \ge 0$ achieves a linear rate.
- If $\alpha > 1$, then $\{\mathbf{x}_t\}_t > 0$ achieves a sublinear rate.
- If $\alpha < 1$, then $\{\mathbf{x}_t\}_t \ge 0$ achieves a superlinear rate.

Trick: Construction Related to Convex L-smooth Function

For a convex and L-smooth function $f(x)$, $\mathbf{x}^*$ is the global minimizer. Construct the function

$$g(\mathbf{x}) = f(\mathbf{x}) - f(\mathbf{x}^*) - \nabla f(\mathbf{x}^*)^\top (\mathbf{x} - \mathbf{x}^*)$$

Then $g(\mathbf{x})$ has a lot of properties:

- $g(\mathbf{x}) \ge 0$ and the equality is achieved when $\mathbf{x} = \mathbf{x}^*$.
- $g(\mathbf{x})$ is still L-smooth and convex.
- $\nabla g(\mathbf{x}) = \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*)$. Thus, $\nabla g(\mathbf{x}^*) = 0$ is the minimizer of $g(\mathbf{x})$

Trick: Fundamental Theorem of Calculus

Goal: For a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, analyze $f(\mathbf{y}) - f(\mathbf{x})$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.
Trick: Consider function $g(t) := f(\mathbf{x} + t(\mathbf{y}-\mathbf{x}))$. $\nabla g(t) = (\mathbf{y}-\mathbf{x})^\top \nabla f(\mathbf{x} + t(\mathbf{y}-\mathbf{x}))$.
We can apply the fundamental theorem of calculus (theorem 2.4) twice for $\Delta := g(1) - g(0)$.

$$\Delta = f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \nabla g(t)dt$$

If f is twice-differentiable, then $\nabla^2 g(t) = (\mathbf{y}-\mathbf{x})^\top \nabla^2 f(\mathbf{x} + t(\mathbf{y}-\mathbf{x}))(\mathbf{y}-\mathbf{x})$.

$$\begin{aligned} \Delta = f(\mathbf{y}) - f(\mathbf{x}) &= \int_0^1 \nabla g(t)dt \\ &= \int_0^1 \left(\int_0^t \nabla^2 g(z)dz + \nabla g(0)\right)dt \end{aligned}$$

2. Convex Functions

Cauchy-Schwarz Inequality

(1) $|\langle \mathbf{u}, \mathbf{v} \rangle|^2 \le \langle \mathbf{u}, \mathbf{u} \rangle \cdot \langle \mathbf{v}, \mathbf{v} \rangle$

(2) $\mathbf{x}^\top \mathbf{y} \le \|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2$

(3) $\mathbf{x}^\top \mathbf{y} \le \|\mathbf{x}\|_\bullet \|\mathbf{y}\|_*$ in general.

(4) $-1 \le \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} \le 1$

(5) Triangle Ineq: $\|\mathbf{u} + \mathbf{v}\| \le \|\mathbf{u}\| + \|\mathbf{v}\|$

(6) $\|\mathbf{a} + \mathbf{b}\|_2^2 \leqslant 2\|\mathbf{a}\|_2^2 + 2\|\mathbf{b}\|_2^2$

(7) Cosine: $\cos\theta_{\mathbf{uv}} = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\|\|\mathbf{v}\|}$

(8) Var, Cov: $\mathrm{Cov}(X, Y)^2 \leqslant \mathrm{Var}(Y)\mathrm{Var}(X)$

(9) Expectation: $\mathbb{E}(XY)^2 \leqslant \mathbb{E}\left(X^2\right)\mathbb{E}\left(Y^2\right)$

Mean Value Theorem

Let $a < b$, $a, b \in \mathbb{R}$, and $h : [a, b] \to \mathbb{R}$ be a continuous func that is differentiable on (a, b). Then there exists $c \in (a, b)$ s.t.

$$h'(c) = \frac{h(b) - h(a)}{b - a}$$

Convex Set
A set $C \subseteq \mathbb{R}^d$ is convex if the line segment between any two points of C lies in C, i.e., if for any $\mathbf{x}, \mathbf{y} \in C$ and any λ with $0 \le \lambda \le 1$, we have

$$\lambda \mathbf{x} + (1-\lambda)\mathbf{y} \in C.$$

Intersections of convex sets are convex.
Convex Functions
A function $f : \mathrm{dom}(f) \to \mathbb{R}$ is convex if (i) $\mathrm{dom}(f)$ is a convex set and (ii) for all $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$, and λ with $0 \le \lambda \le 1$, we have

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \le \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$$

- The graph of a function $f : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\{(\mathbf{x}, f(\mathbf{x})) \mid \mathbf{x} \in \mathrm{dom}(f)\}$$

- The epigraph of a function $f : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\mathrm{epi}(f) := \left\{(\mathbf{x}, \alpha) \in \mathbb{R}^{d+1} \mid \mathbf{x} \in \mathrm{dom}(f), \alpha \ge f(\mathbf{x})\right\},$$

- f is a convex function if and only if $\mathrm{epi}(f)$ is a convex set.

Examples of convex functions
- Affine, Square, Exponential
- Norm: Every norm is convex.

Convex Functions are Continuous
Let f be convex and suppose that $\mathrm{dom}(f) \subseteq \mathbb{R}^d$ is open. Then f is continuous.

Jensen's Inequality
Let f be convex, $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbf{dom}(f), \lambda_1, \ldots, \lambda_m \in \mathbb{R}_+$such that $\sum_{i=1}^m \lambda_i = 1$. Then

$$f\left(\sum_{i=1}^m \lambda_i \mathbf{x}_i\right) \le \sum_{i=1}^m \lambda_i f(\mathbf{x}_i)$$

- Expectation: If X is a random variable and φ is a convex function, then

$$\varphi(\mathbb{E}[X]) \le \mathbb{E}[\varphi(X)]$$

Differentiable Functions
Let $f : \mathrm{dom}(f) \to \mathbb{R}^m$ where $\mathrm{dom}(f) \subseteq \mathbb{R}^d$ is open. f is called differentiable at $\mathbf{x} \in \mathrm{dom}(f)$ if there exists an $(m \times d)$-matrix $\mathbf{A}$ and an error function $r : \mathbb{R}^d \to \mathbb{R}^m$ defined around $\mathbf{0} \in \mathbb{R}^d$ such that for all $\mathbf{y}$ in some neighborhood of $\mathbf{x}$,

$$f(\mathbf{y}) = f(\mathbf{x}) + \mathbf{A}(\mathbf{y} - \mathbf{x}) + r(\mathbf{y} - \mathbf{x})$$

where

$$\lim_{\mathbf{v} \to \mathbf{0}} \frac{\|r(\mathbf{v})\|}{\|\mathbf{v}\|} = \mathbf{0}. \quad \text{(Error } r \text{ is sublinear)}$$

- $\mathbf{A}$ is unique and called the differential or Jacobian matrix of f at $\mathbf{x}$.
- Graph of the affine function $f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x})$ is a tangent hyperplane to the graph of f at $(\mathbf{x}, f(\mathbf{x}))$.

Lemma 2.15 First-order Characterization of Convexity
f is convex if and only if $\mathrm{dom}(f)$ **is convex** and $f(\mathbf{y}) \ge f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x})$ holds for all $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$.

Lemma 2.17 Second-order Characterization of Convexity
f is convex if and only if $\mathrm{dom}(f)$ **is convex**, and for all $\mathbf{x} \in \mathrm{dom}(f)$, we have $\nabla^2 f(\mathbf{x}) \succeq 0$

Lemma 2.16 Monotonicity of the Gradient
Suppose that $\mathrm{dom}(f)$ is open and that f is differentiable. Then f is convex **iff** $\mathrm{dom}(f)$ is convex and

$$(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))^\top(\mathbf{y} - \mathbf{x}) \ge 0$$

holds for all $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$.
The inequality in monotonicity of the gradient is strict unless $\mathbf{x} = \mathbf{y}$ or $f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) = \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$ for all $\lambda \in (0, 1)$.

Lemma 2.18 Operations that Preserve Convexity
(1) Let $f_1, f_2, \ldots, f_m$ be **convex** functions, $\lambda_1, \lambda_2, \ldots, \lambda_m \in \mathbb{R}_+$. Then $f := \max_{i=1}^m f_i$ as well as $f := \sum_{i=1}^m \lambda_i f_i$ are convex on $\mathrm{dom}(f) := \bigcap_{i=1}^m \mathrm{dom}(f_i)$.

(2) Let f be a **convex** function with $\mathrm{dom}(f) \subseteq \mathbb{R}^d$, $g : \mathbb{R}^m \to \mathbb{R}^d$, $g(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$, for some matrix $\mathbf{A} \in \mathbb{R}^{d \times m}$ and vector $\mathbf{b} \in \mathbb{R}^d$. Then the function $f \circ g$ is convex on $\mathrm{dom}(f \circ g) := \{\mathbf{x} \in \mathbb{R}^m : g(\mathbf{x}) \in \mathrm{dom}(f)\}$.

Local & Global Minima
A local minimum of $f : \mathrm{dom}(f) \to \mathbb{R}$ is a point $\mathbf{x}$ such that there exists $\varepsilon > 0$ with

$$f(\mathbf{x}) \le f(\mathbf{y}) \quad \forall \mathbf{y} \in \mathbf{dom}(f) \text{ satisfying } \|\mathbf{y} - \mathbf{x}\| < \varepsilon$$

Meaning: in some small neighborhood, $\mathbf{x}$ is the best point.

Lemma 2.20
Let $\mathbf{x}^\star$ be a local minimum of a convex function $f : \mathrm{dom}(f) \to \mathbb{R}$. Then $\mathbf{x}^\star$ is a global minimum, meaning that $f(\mathbf{x}^\star) \le f(\mathbf{y}) \quad \forall \mathbf{y} \in \mathrm{dom}(f)$.

Lemma 2.22
Suppose that $f : \mathrm{dom}(f) \to \mathbb{R}$ is differentiable over an open domain $\mathrm{dom}(f) \subseteq \mathbb{R}^d$. Let $\mathbf{x} \in \mathrm{dom}(f)$. If $\mathbf{x}$ is a global minimum then $\nabla f(\mathbf{x}) = \mathbf{0}$ (a critical point).

Lemma 2.21
For convex func, the converse of Lemma 2.22 is also true: If $\nabla f(\mathbf{x}) = \mathbf{0}$, then $\mathbf{x}$ is a global minimum.

Strictly Convex Functions
A function $f : \mathrm{dom}(f) \to \mathbb{R}$ is strictly convex if (i) $\mathrm{dom}(f)$ is convex and (ii) for all $\mathbf{x} \ne \mathbf{y} \in \mathrm{dom}(f)$ and all $\lambda \in (0, 1)$, we have

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$$

Lemma 2.25
Strictly convex func has **at most one** global min.

Constrained Minimization
A point $\mathbf{x} \in X$ is a minimizer of f over X if

$$f(\mathbf{x}) \le f(\mathbf{y}) \quad \forall \mathbf{y} \in X.$$

Lemma 2.27 Optimality Condition
Suppose that $f : \mathrm{dom}(f) \to \mathbb{R}$ is convex and differentiable over an open domain $\mathrm{dom}(f) \subseteq \mathbb{R}^d$, and let $X \subseteq \mathrm{dom}(f)$ be a convex set. Point $\mathbf{x}^\star \in X$ is a minimizer of f over X if and only if

$$\nabla f\left(\mathbf{x}^\star\right)^\top \left(\mathbf{x} - \mathbf{x}^\star\right) \ge 0 \quad \forall \mathbf{x} \in X$$

Sublevel
$f : \mathbb{R}^d \to \mathbb{R}, \alpha \in \mathbb{R}$. The set $f^{\le \alpha} := \left\{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) \le \alpha\right\}$ is the α-sublevel set of f.

Thm 2.29 (Weierstrass Theorem)
Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function, and suppose there is a nonempty and bounded sublevel set $f^{\le \alpha}$. Then f has a global minimum.

Optimization Problem in Standard Forms

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

Domain $\mathcal{D} = \left\{\cap_{i=0}^m \mathrm{dom}(f_i)\right\} \cap \left\{\cap_{i=1}^p \mathrm{dom}(h_i)\right\}$
Convex program: All f_i are convex functions, and all h_i are affine functions with domain $\mathbb{R}^d$.

Lagrangian
Given an optimization problem in standard form, its Lagrangian is the func $L : \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ given by

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i h_i(\mathbf{x})$$

The λ_i, ν_i are called Lagrange multipliers. The Lagrange dual function is the function $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R} \cup \{-\infty\}$ defined by

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x} \in \mathcal{D}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}).$$

Lemma 2.45 Weak Lagrange Duality
Short Ver.: Lagrange dual function values are lower bounds on primal function values $f_0(\mathbf{x})$.
Long Ver.: Let $\mathbf{x}$ be a feasible solution, meaning that $f_i(\mathbf{x}) \le 0$ for $i = 1, \ldots, m$ and $h_i(\mathbf{x}) = 0$ for $i = 1, \ldots, p$. Let g be the Lagrange dual function of and $\boldsymbol{\lambda} \in \mathbb{R}^m, \boldsymbol{\nu} \in \mathbb{R}^p$ such that $\boldsymbol{\lambda} \ge \mathbf{0}$. Then

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \le f_0(\mathbf{x})$$

Choose $\boldsymbol{\lambda} \ge \mathbf{0}$ and $\boldsymbol{\nu}$ such that $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is maximized! By weak duality, the supremum value of the Lagrange dual is a lower bound for the infimum value of the primal problem.

Thm 2.47 Strong Lagrange Duality
Suppose that a convex program has a feasible solution $\tilde{\mathbf{x}}$ that in addition satisfies $f_i(\tilde{\mathbf{x}}) < 0, i = 1, \ldots, m$ (a Slater point). Then the infimum value of the primal equals the supremum value of its Lagrange dual. Moreover, if this value is finite, it is attained by a feasible solution of the dual.
Convex programming with Slater point and finite value: $\inf f_0(\mathbf{x}) = \max g(\boldsymbol{\lambda}, \boldsymbol{\nu})$.

Zero Duality Gap
Let $\tilde{\mathbf{x}}$ be feasible for the primal and $(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$ feasible for the Lagrange dual. The primal and dual solutions $\tilde{\mathbf{x}}$ and $(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$ are said to have zero duality gap if $f_0(\tilde{\mathbf{x}}) = g(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$.

Lemma 2.49 Complementary Slackness
If $\tilde{\mathbf{x}}$ and $(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$ have zero duality gap, then

$$\tilde{\lambda}_i f_i(\tilde{\mathbf{x}}) = 0, \quad i = 1, \ldots, m.$$

Lemma 2.50 Vanishing Lagrangian Gradient
If $\tilde{\mathbf{x}}$ and $(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$ have zero duality gap, and if all f_i and h_i are differentiable, then

$$\nabla f_0(\tilde{\mathbf{x}}) + \sum_{i=1}^m \tilde{\lambda}_i \nabla f_i(\tilde{\mathbf{x}}) + \sum_{i=1}^p \tilde{\nu}_i \nabla h_i(\tilde{\mathbf{x}}) = \mathbf{0}$$

KKT Conditions
- primal and dual feasibilty
- complementary slackness
- vanishing Lagrangian gradient

$$\begin{aligned} f_i(\tilde{x}) &\le 0, \quad i = 1, \ldots, m \\ h_i(\tilde{x}) &= 0, \quad i = 1, \ldots, p \\ \tilde{\lambda}_i &\ge 0, \quad i = 1, \ldots, m \\ \tilde{\lambda}_i f_i(\tilde{x}) &= 0, \quad i = 1, \ldots, m \end{aligned}$$

$$\nabla f_0(\tilde{x}) + \sum_{i=1}^m \tilde{\lambda}_i \nabla f_i(\tilde{x}) + \sum_{i=1}^p \tilde{\nu}_i \nabla h_i(\tilde{x}) = 0$$

Thm 2.52
Suppose that all f_i and h_i are differentiable, all f_i are convex, all h_i are affine. Let $\tilde{\mathbf{x}}$ and $(\tilde{\lambda}, \tilde{\nu})$ be such that the KKT conditions hold. Then $\tilde{\mathbf{x}}$ and $(\tilde{\lambda}, \tilde{\nu})$ have zero duality gap and hence are primal and dual optimal solutions.

3. Gradient Descent

Idea

- Iterative Algorithm: Choose $\mathbf{x}_0 \in \mathbb{R}^d$. $\mathbf{x}_{t+1} := \mathbf{x}_t + \mathbf{v}_t$ for times $t = 0, 1, \ldots$, and steps $\mathbf{v}_t \in \mathbb{R}^d$.
- Gradient descent: Choose $\mathbf{x}_0 \in \mathbb{R}^d$ $\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t)$ for times $t = 0, 1, \ldots$, and stepsize $\gamma \ge 0$.
- Abbreviate: $\mathbf{g}_t := \nabla f(\mathbf{x}_t)$ (gradient descent: $\mathbf{g}_t = (\mathbf{x}_t - \mathbf{x}_{t+1})/\gamma$)

Vanilla Analysis

$\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) = \frac{\gamma}{2}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\left(\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right)$

Upper bound for the average error $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ over the first T iterations:

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) \le \frac{1}{T}\left(\frac{\gamma}{2}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2\right)$$

Lipschitz Convex Func: $\mathcal{O}(1/\varepsilon^2)$ Steps

Assume that all gradients of f are bounded in norm. Equivalent to f being Lipschitz (Thm 2.9).

Thm 3.1
Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum $\mathbf{x}^\star$; furthermore, suppose that $\|\mathbf{x}_0 - \mathbf{x}^\star\| \le R$ and $\|\nabla f(\mathbf{x})\| \le B$ for all $\mathbf{x}$. Choosing the stepsize $\gamma := \frac{R}{B\sqrt{T}}$ gradient descent yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) \le \frac{RB}{\sqrt{T}}$$

Average error $\le \frac{RB}{\sqrt{T}} \le \varepsilon \Rightarrow T \ge \frac{R^2B^2}{\varepsilon^2}$

Smooth Functions

Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be differentiable, $X \subseteq \mathbf{dom}(f)$ convex, $L \in \mathbb{R}_+$. f is called smooth (with parameter L) over X if

$$f(\mathbf{y}) \le f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in X.$$

- Does not require convexity.
- In general, quadratic functions are smooth.
- Affine functions are smooth with param 0.

Lemma 3.3
Smoothness of $f(\mathbf{x})$ = convexity of $\frac{L}{2}\mathbf{x}^\top\mathbf{x} - f(\mathbf{x})$

Lemma 3.4
Let $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{Q}\mathbf{x} + \mathbf{b}^\top\mathbf{x} + c$, where $\mathbf{Q}$ is a symmetric $(d \times d)$ matrix, $\mathbf{b} \in \mathbb{R}^d, c \in \mathbb{R}$. Then f is smooth with parameter $2\|\mathbf{Q}\|_2$.

Lemma 3.5
Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. The following two statements are equivalent.

(1) f is smooth with parameter L.

(2) $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \le L\|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

Lemma 3.6 Operations that preserve smoothness

(1) Let $f_1, f_2, \ldots, f_m$ be functions that are smooth with parameters $L_1, L_2, \ldots, L_m$, and let $\lambda_1, \lambda_2, \ldots, \lambda_m \in \mathbb{R}_+$. Then the function $f := \sum_{i=1}^m \lambda_i f_i$ is smooth with parameter $\sum_{i=1}^m \lambda_i L_i$.

(2) Let f be smooth with parameter L, and let $g(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$, for $\mathbf{A} \in \mathbb{R}^{d \times m}$ and $\mathbf{b} \in \mathbb{R}^d$. Then the function $f \circ g$ is smooth with parameter $L\|\mathbf{A}\|_2^2$.

Lemma 3.7 Sufficient Decrease
Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and smooth with parameter L. With stepsize $\gamma := \frac{1}{L}$ gradient descent satisfies

$$f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2, \quad t \ge 0$$

This doesn't require convexity.

Corollary of Sufficient Decrease Lemma
Let g be a L-smooth function and $\mathbf{x}^*$ be a minimizer of g. Then for any $\mathbf{x} \in \mathbf{dom}(g)$, we have

$$g(\mathbf{x}) - g(\mathbf{x}^*) \ge \frac{1}{2L}\|\nabla g(\mathbf{x})\|_2^2.$$

(**Proof see Hw6 Ex1**)

Thm 3.8 Smooth Convex Func: $\mathcal{O}(1/\varepsilon)$ Steps

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum $\mathbf{x}^\star$; furthermore, suppose that f is smooth with parameter L. Choosing stepsize $\gamma := \frac{1}{L}$ gradient descent yields $f(\mathbf{x}_T) - f(\mathbf{x}^\star) \le \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$

Let $R^2 := \|\mathbf{x}_0 - \mathbf{x}^\star\|^2$, error $\le \frac{L}{2T}R^2 \le \varepsilon \Rightarrow T \ge \frac{R^2L}{2\varepsilon}$

Nesterov's Accelerated Gradient Descent (AGD): $\mathcal{O}(1/\sqrt{\varepsilon})$ Steps

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex, differentiable, and smooth with parameter L. Choose $\mathbf{z}_0 = \mathbf{y}_0 = \mathbf{x}_0$ arbitrary. For $t \ge 0$, set

$$\mathbf{y}_{t+1} := \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)$$
$$\mathbf{z}_{t+1} := \mathbf{z}_t - \frac{t+1}{2L}\nabla f(\mathbf{x}_t)$$
$$\mathbf{x}_{t+1} := \frac{t+1}{t+3}\mathbf{y}_{t+1} + \frac{2}{t+3}\mathbf{z}_{t+1}$$

Thm 3.9 Error Bound of AGD
Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum $\mathbf{x}^\star$; furthermore, suppose that f is smooth with parameter L. Accelerated gradient descent yields

$$f(\mathbf{y}_T) - f(\mathbf{x}^\star) \le \frac{2L\|\mathbf{z}_0 - \mathbf{x}^\star\|^2}{T(T+1)}, \quad T > 0.$$

Potential Function of AGD
Define the potential as

$$\Phi(t) := t(t+1)\left(f(\mathbf{y}_t) - f(\mathbf{x}^\star)\right) + 2L\|\mathbf{z}_t - \mathbf{x}^\star\|^2.$$

We can show that $\Phi(t+1) \le \Phi(t)$ for every t. Rewriting $\Phi(T) \le \Phi(0)$, we can claim the error bound in Thm 3.9.
(**Proof see Handout03 Pages 29-30**)

Strongly Convex Func
Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be a convex and differentiable function, $X \subseteq \mathbf{dom}(f)$ convex and $\mu \in \mathbb{R}_+, \mu > 0$. Function f is called strongly convex (with parameter μ) over X if

$$f(\mathbf{y}) \ge f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in X.$$

Lemma 3.11
Suppose that $\mathbf{dom}(f)$ is open and convex, and that $f : \mathbf{dom}(f) \to \mathbb{R}$ is differentiable. Let $\mu \in \mathbb{R}_+$. Then the following two statements are equivalent. (i) f is strongly convex with parameter μ. (ii) g defined by $g(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2}\mathbf{x}^\top\mathbf{x}$ is convex over $\mathbf{dom}(g) := \mathbf{dom}(f)$.

- f is m-strongly convex iff $f''(\mathbf{x}) \ge m > 0$ for all $\mathbf{x}$.

Lemma 3.12
If $f : \mathbb{R}^d \to \mathbb{R}$ is strongly cvx with param $\mu > 0$, then f is strictly cvx and has a unique global min.

Thm 3.14 Smooth and strongly convex func: $\mathcal{O}(\log(1/\varepsilon))$ Steps

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum $\mathbf{x}^\star$; suppose that f is smooth with parameter L and strongly convex with parameter $\mu > 0$. Choosing $\gamma := \frac{1}{L}$, gradient descent with arbitrary $\mathbf{x}_0$ satisfies the following two properties.
(i) Squared distances to $\mathbf{x}^\star$ are geometrically decreasing:

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \le \left(1 - \frac{\mu}{L}\right)\|\mathbf{x}_t - \mathbf{x}^\star\|^2, \quad t \ge 0$$
$$\|\mathbf{x}_T - \mathbf{x}^\star\|^2 \le \left(1 - \frac{\mu}{L}\right)^T\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$$

(ii) The absolute error after T iterations is exponentially small in T :

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \le \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

Let $R^2 := \|\mathbf{x}_0 - \mathbf{x}^\star\|^2$. Then error $\le \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T R^2 \le \varepsilon \Rightarrow T \ge \frac{L}{\mu}\ln\left(\frac{R^2L}{2\varepsilon}\right)$

Summary: Lipschitz Continuous Gradient (L-smoothness)

A differentiable function f is said to have an L-Lipschitz continuous gradient if for some $L > 0$

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \le L\|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y}$$

Note: The definition doesn't assume convexity of f.

(0) $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \le L\|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y}$.

(1) $g(\mathbf{x}) = \frac{L}{2}\mathbf{x}^\top\mathbf{x} - f(\mathbf{x})$ is convex, $\forall \mathbf{x}$

(2) $f(\mathbf{y}) \le f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2, \forall \mathbf{x}, \mathbf{y}$.

(3) $(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top(\mathbf{x} - \mathbf{y}) \le L\|\mathbf{x} - \mathbf{y}\|^2, \forall \mathbf{x}, \mathbf{y}$

(4) $f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \ge \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y}) - \frac{\alpha(1-\alpha)L}{2}\|\mathbf{x} - \mathbf{y}\|^2, \forall \mathbf{x}, \mathbf{y}$ and $\alpha \in [0, 1]$

(5) $f(\mathbf{y}) \ge f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{1}{2L}\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|^2, \forall \mathbf{x}, \mathbf{y}$.

(6) $(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top(\mathbf{x} - \mathbf{y}) \ge \frac{1}{L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2, \forall \mathbf{x}, \mathbf{y}$

(7) $f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \le \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y}) - \frac{\alpha(1-\alpha)}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2, \forall \mathbf{x}, \mathbf{y}$ and $\alpha \in [0, 1]$

Note: We assume that the domain for f and g are both $\mathbb{R}^n$, and hence convex set.
For a function f with a Lipschitz continuous gradient over $\mathbb{R}^n$, the following implications hold:

$$[5] \equiv [7] \to [6] \to [0] \to [1] \equiv [2] \equiv [3] \equiv [4]$$

If we further assume that f is convex, then we have all the conditions $[0] - [7]$ are equivalent.

Summary: Strong Convexity

A differentiable function f is strongly convex if $f(\mathbf{y}) \ge f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2$ for some $\mu > 0$ and all x, y Note: Strong convexity doesn't necessarily require the function to be differentiable, and the gradient is replaced by the sub-gradient when the function is non-smooth.

Equivalent Conditions of Strong Convexity

The following conditions are all equivalent to the condition that a differentiable function f is strongly-convex with constant $\mu > 0$.

(i) $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|^2, \forall \mathbf{x}, \mathbf{y}$.

(ii) $g(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|^2$ is convex, $\forall x$.

(iii) $(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top(\mathbf{x}-\mathbf{y}) \geq \mu\|\mathbf{x}-\mathbf{y}\|^2, \forall \mathbf{x}, \mathbf{y}$.

(iv) $f(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}) - \frac{\alpha(1-\alpha)\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2, \alpha \in [0,1]$.

For a continuously differentiable function f, the following conditions are all implied by strong convexity (SC) condition.

(a) $\frac{1}{2}\|\nabla f(\mathbf{x})\|^2 \geq \mu(f(\mathbf{x}) - f^*), \forall \mathbf{x}$.

(b) $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \geq \mu\|\mathbf{x}-\mathbf{y}\| \forall \mathbf{x}, \mathbf{y}$.

(c) $f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) + \frac{1}{2\mu}\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|^2, \forall \mathbf{x}, \mathbf{y}$.

(d) $(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top(\mathbf{x}-\mathbf{y}) \leq \frac{1}{\mu}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2, \forall \mathbf{x}, \mathbf{y}$

Additivity of Strongly Convex Functions

- Assume that real functions f is a-strongly convex and g is b-strongly convex. Then the sum $f+g$ is also strongly convex, with parameter $a+b$. (**Proof see GA2 Solution6**)
- Let $h = f + g$ where f is strongly convex with param μ and g is convex, then h is strongly convex with param μ.

4a. Projected Gradient Descent

Constrained Optimization Problem

$$\text{minimize } f(\mathbf{x})$$
$$\text{subject to } \mathbf{x} \in X, X \subsetneq \mathbb{R}^d \text{ (closed convex set)}$$

Projected Gradient Descent

Choose $\mathbf{x}_0 \in \mathbb{R}^d$.

$$\mathbf{y}_{t+1} := \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t)$$
$$\mathbf{x}_{t+1} := \Pi_X(\mathbf{y}_{t+1}) := \underset{\mathbf{x}\in X}{\operatorname{argmin}} \|\mathbf{x}-\mathbf{y}_{t+1}\|^2$$

for **times** $t = 0, 1, \ldots$, and **stepsize** $\gamma \geq 0$.

- When $\nabla f(\mathbf{x}_t) \neq \mathbf{0}$ but $\mathbf{x}_{t+1} = \mathbf{x}_t$ (convex f and X), **we have reached an optimal solution**: the gradient is orthogonal to a hyperplane through $\mathbf{x}_t$ that has all feasible solutions on one side.
- $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t)$ (**Proof see Hw4 Ex1**)
- $\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\| \leq \|\mathbf{y}_{t+1} - \mathbf{x}_t\| = \gamma\|\nabla f(\mathbf{x}_t)\|$ (**Proof see Hw4 Ex1**)

Optimality Condition for Projected Gradient Descent

Using projected gradient descent, $\mathbf{x}^*$ is an optimal solution to the constrained optimization problem **iff** $\mathbf{x}^* = \Pi_X(\mathbf{x}^* - \gamma\nabla f(\mathbf{x}^*))$

(**Proof see ODS Exam FS20 Assignment 2**)

Fact 4a.1 Properties of Projection

Let $X \subseteq \mathbb{R}^d$ closed and convex, $\mathbf{x} \in X, \mathbf{y} \in \mathbb{R}^d$. Then

(i) $(\mathbf{x} - \Pi_X(\mathbf{y}))^\top(\mathbf{y} - \Pi_X(\mathbf{y})) \leq 0$.

(ii) $\|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{y} - \Pi_X(\mathbf{y})\|^2 \leq \|\mathbf{x}-\mathbf{y}\|^2$.

Num of Steps for PGD

The **same number of steps as GD** over $\mathbb{R}^d$!

- Lipschitz convex functions over $X: \mathcal{O}(1/\varepsilon^2)$ steps
- Smooth convex functions over $X: \mathcal{O}(1/\varepsilon)$ steps
- Smooth and strongly convex functions over $X: \mathcal{O}(\log(1/\varepsilon))$ steps

Lemma 4a.3 Projected Sufficient Decrease

Let $f: \mathrm{dom}(f) \to \mathbb{R}$ be differentiable and smooth with parameter L over a closed and convex set $X \subseteq \mathrm{dom}(f)$. Choosing stepsize $\gamma := \frac{1}{L}$, projected gradient descent with arbitrary $\mathbf{x}_0 \in X$ satisfies, for $t \geq 0$,

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2$$

Thm 4a.4 Smooth Convex Func over $X: \mathcal{O}(1/\varepsilon)$ Steps

Let $f: \mathrm{dom}(f) \to \mathbb{R}$ be convex and differentiable. Let $X \subseteq \mathrm{dom}(f)$ be a closed convex set, and assume that there is a minimizer $\mathbf{x}^\star$ of f over X; furthermore, suppose that f is smooth over X with parameter L. Choosing stepsize $\gamma := \frac{1}{L}$, projected gradient descent yields

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t)$$

and

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2T}\left\|\mathbf{x}_0 - \mathbf{x}^\star\right\|^2, \quad T > 0.$$

Exactly the **same bound** as in the unconstrained case!

4b. Coordinate Descent

PL Inequality

Let $f: \mathbb{R}^d \to \mathbb{R}$ be a differentiable function with a global minimum $\mathbf{x}^\star$. We say that f satisfies the **PL inequality** if the following holds for some $\mu > 0$:

$$\frac{1}{2}\|\nabla f(\mathbf{x})\|^2 \geq \mu\left(f(\mathbf{x}) - f(\mathbf{x}^\star)\right), \quad \forall \mathbf{x} \in \mathbb{R}^d$$

Direct consequence: $\nabla f(\mathbf{x}) = \mathbf{0} \Rightarrow \mathbf{x}$ is a global min.

Lemma 4b.2 Strong convexity $\Rightarrow$ PL inequality

Let $f: \mathbb{R}^d \to \mathbb{R}$ be differentiable and strongly convex with parameter $\mu > 0$ (in particular, a global minimum $x^\star$ exists by Lemma 3.12). Then f satisfies the PL inequality for the same μ.

Thm 4b.3 GD on Smooth Func with PL Ineq

Let $f: \mathbb{R}^d \to \mathbb{R}$ be differentiable with a global min $\mathbf{x}^\star$. Suppose that f is smooth with param L and satisfies the PL ineq with param $\mu > 0$. Choosing stepsize $\gamma = 1/L$, GD with arbitrary $\mathbf{x}_0$ satisfies

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \left(1 - \frac{\mu}{L}\right)^T (f\mathbf{x}_0) - f(\mathbf{x}^\star)\Big), \quad T > 0$$

Coordinate-wise Smoothness

Let $f: \mathbb{R}^d \to \mathbb{R}$ be differentiable, and $\mathcal{L} = (L_1, L_2, \ldots, L_d) \in \mathbb{R}^d_+$. Func f is called **coordinate-wise smooth** (with param $\mathcal{L}$) if for every coordinate $i = 1, 2, \ldots, d$,

$$f(\mathbf{x} + \lambda\mathbf{e}_i) \leq f(\mathbf{x}) + \lambda\nabla_i f(\mathbf{x}) + \frac{L_i}{2}\lambda^2 \quad \forall \mathbf{x} \in \mathbb{R}^d, \lambda \in \mathbb{R},$$

- If $L_i = L$ for all i, f is said to be coordinate-wise smooth with param L.
- If f is smooth with param L, then f is coordinate-wise smooth with param L.

Coordinate Descent Algo

In Iteration t:

(i) Choose some $i \in [d]$

(ii) $\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_i \nabla_i f(\mathbf{x}_t)\mathbf{e}_i$

- $\nabla_i f(\mathbf{x}_t)$ is the i-th entry of the gradient (i-th partial derivate).
- $\mathbf{e}_i$ is the i-th unit vector, so only the i-th coordinate of $\mathbf{x}_t$ is updated.
- γ_i is the stepsize for coordinate i.

Lemma 4b.5 Coordinate-wise Sufficient Decrease

Let $f: \mathbb{R}^d \to \mathbb{R}$ be differentiable and coordinate-wise smooth with parameter $\mathcal{L} = (L_1, L_2, \ldots, L_d)$. With active coordinate i in iteration t and stepsize $\gamma_i = \frac{1}{L_i}$, coordinate descent satisfies

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L_i}\|\nabla_i f(\mathbf{x}_t)\|^2$$

Randomized Coordinate Descent

(i) sample $i \in [d]$ uniformly at random

(ii) $\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_i \nabla_i f(\mathbf{x}_t)\mathbf{e}_i$

Thm 4b.6 Randomized Coordinate Descent: Smooth Func, PL inequality

Let $f: \mathbb{R}^d \to \mathbb{R}$ be differentiable with a global minimum $\mathbf{x}^\star$. Suppose that f is coordinate-wise smooth with parameter L and satisfies the PL inequality with parameter $\mu > 0$. Choosing stepsize $\gamma_i = 1/L$ for all coordinates, randomized coordinate descent with arbitrary $\mathbf{x}_0$ satisfies

$$\mathbb{E}\left[f(\mathbf{x}_T) - f(\mathbf{x}^\star)\right] \leq \left(1 - \frac{\mu}{dL}\right)^T \left(f(\mathbf{x}_0) - f(\mathbf{x}^\star)\right)$$

Importance Sampling

Improves over uniform sampling when coordinate-wise smoothness parameters L_i differ.

(i) sample $i \in [d]$ with probability $\frac{L_i}{\sum_{j=d}^d L_j}$

(ii) $\mathbf{x}_{t+1} := \mathbf{x}_t - \frac{1}{L_i}\nabla_i f(\mathbf{x}_t)\mathbf{e}_i$

Thm 4b.7 Convergence of Importance Sampling

Let $f: \mathbb{R}^d \to \mathbb{R}$ be differentiable with a global min $\mathbf{x}^\star$, coordinate-wise smooth with param $\mathcal{L} = (L_1, L_2, \ldots, L_d)$, and satisfying the PL ineq with param $\mu > 0$. Let $\bar{L} = \frac{1}{d}\sum_{i=1}^d L_i$. Then coordinate descent with **importance sampling** and arbitrary $\mathbf{x}_0$ satisfies

$$\mathbb{E}\left[f(\mathbf{x}_T) - f(\mathbf{x}^\star)\right] \leq \left(1 - \frac{\mu}{d\bar{L}}\right)^T \left(f(\mathbf{x}_0) - f(\mathbf{x}^\star)\right)$$

Corollary 4b.8

Same number of iterations as randomized coordinate descent.

Strong convexity wrt ℓ_1-norm

- Measure strong convexity wrt ℓ_1-norm instead of ℓ_2-norm:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) + \frac{\mu_1}{2}\|\mathbf{y}-\mathbf{x}\|_1^2, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$$

- Then f is also strongly convex with $\mu = \mu_1$ in the usual sense.

  Proof: $\|\mathbf{y}-\mathbf{x}\|_1 \geq \|\mathbf{y}-\mathbf{x}\|_2$
- If f is strongly convex with μ in the usual sense, then f is strongly convex with $\mu_1 = \mu/d$ w.r.t. ℓ_1-norm.

  Proof: $\|\mathbf{y}-\mathbf{x}\| \geq \|\mathbf{y}-\mathbf{x}\|_1/\sqrt{d}$
- $\mu \geq \mu_1 \geq \mu/d$
- If $\mu_1 > \mu/d$, we can speed up steepest coordinate descent.

Lemma 4b.9 Strong convexity w.r.t. ℓ_1-norm $\Rightarrow$ PL inequality w.r.t. ℓ_∞-norm

Let $f: \mathbb{R}^d \to \mathbb{R}$ be differentiable and strongly convex with parameter $\mu_1 > 0$ w.r.t. ℓ_1-norm. (In particular, f is μ_1-strongly convex w.r.t. Euclidean norm, so a global minimum $x^\star$ exists by Lemma 3.12). Then f satisfies the PL inequality w.r.t. ℓ_∞-norm with the same μ_1:

$$\frac{1}{2}\|\nabla f(\mathbf{x})\|_\infty^2 \geq \mu_1\left(f(\mathbf{x}) - f(\mathbf{x}^\star)\right), \quad \forall \mathbf{x} \in \mathbb{R}^d$$

Thm 4b.10 Steeper (than steepest) coordinate descent

Let $f:\mathbb{R}^d \to \mathbb{R}$ be differentiable with a global minimum $\mathbf{x}^\star$. Suppose that f is coordinate-wise smooth with parameter L and satisfies the PL inequality w.r.t. ℓ_∞-norm with parameter $\mu_1 > 0$. Choosing stepsize $\gamma_i = 1/L$, steepest coordinate descent with arbitrary $\mathbf{x}_0$ satisfies

$$f(\mathbf{x}_T) - f\left(\mathbf{x}^\star\right) \le \left(1 - \frac{\mu_1}{L}\right)^T \left(f(\mathbf{x}_0) - f\left(\mathbf{x}^\star\right)\right), \quad T > 0$$

- Normal steepest coordinate descent: $\left(1 - \frac{\mu}{dL}\right)$.
- Worst case: $\mu_1 = \mu/d$, no speedup.
- Best case: $\mu_1 = \mu$, speedup by a factor of d.

Greedy Coordinate Descent

Make the step that maximizes the progress in the chosen coordinate!

(i) choose $i \in [d]$

(ii) $\mathbf{x}_{t+1} := \operatorname{argmin}_{\lambda \in \mathbb{R}} f(\mathbf{x}_t + \lambda \mathbf{e}_i)$

This requires to perform a line search.

Thm 4b.11

Let $f:\mathbb{R}^d \to \mathbb{R}$ be of the form

$$f(\mathbf{x}) := g(\mathbf{x}) + h(\mathbf{x}) \quad \text{with } h(\mathbf{x}) = \sum_i h_i(x_i), \quad \mathbf{x} \in \mathbb{R}^d$$

with g convex and differentiable, and the h_i convex.

Let $\mathbf{x} \in \mathbb{R}^d$ be a point such that greedy coordinate descent cannot make progress in any coordinate. Then $\mathbf{x}$ is a global minimum of f.

A function h as in the theorem is called separable.

Popular examples: regularizers $h(\mathbf{x}) = \|\mathbf{x}\|_1$ and $h(\mathbf{x}) = \|\mathbf{x}\|^2$.

Summary

Algo	Norm	Smooth	Bound	Result
Rand	ℓ_2	L	$1-\frac{\mu}{dL}$	Thm 4b.6
IS	ℓ_2	$L_1\ldots,L_d$	$1-\frac{\mu}{d\bar{L}}$	Thm 4b.7
Steepest	ℓ_2	L	$1-\frac{\mu}{dL}$	Coro 4b.8
Steeper	ℓ_1	L	$1-\frac{\mu_1}{L}$	Thm 4b.10

In the worst case, nothing is gained over gradient descent, and Steepest may even lose.

In the best case, Importance sampling and Steeper (than Steepest) may be up to d times faster than gradient descent.

5. Subgradient Methods

Loss Func

(1) 0-1 Loss: $f(s) = \begin{cases} 1, & s < 0 \\ 0, & s \ge 0 \end{cases}$

(2) Hinge losss: $f(s) = \max(0, 1-s)$

(3) Squared loss: $f(s) = (s-1)^2$

(4) Exponential loss: $f(s) = e^{-s}$

(5) Logistic loss: $f(s) = \log(1 + e^{-s})$

(1) non-convex, (2)-(5) convex. (1)(2)(4) non-smooth, (3)(5) smooth.

Subgradients

$\mathbf{g} \in \mathbb{R}^d$ is a subgradient (not always unique) of f at $\mathbf{x}$ if

$$f(\mathbf{y}) \ge f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{y} - \mathbf{x}) \quad \text{for all } \mathbf{y} \in \operatorname{dom}(f)$$

$\partial f(\mathbf{x}) \subseteq \mathbb{R}^d$ is the subdifferential, the set of subgradients of f at $\mathbf{x}$.

Lemma 5.2 Subgradients of Differentiable Func

If f is differentiable at $\mathbf{x} \in \mathbf{dom}(f)$, then $\partial f(\mathbf{x}) \subseteq \{\nabla f(\mathbf{x})\}$.

i.e., either exactly one subgradient $\nabla f(\mathbf{x})$, or no subgradient at all.

Lemma 5.3 Subgradient Characterization of Convexity

(i) If f is convex, then $\partial f(\mathbf{x}) \neq \emptyset$ for all $\mathbf{x}$ in the (relative) interior of $\operatorname{dom}(f)$.

(ii) If $\operatorname{dom}(f)$ is convex and $\partial f(\mathbf{x}) \neq \emptyset$ for all $\mathbf{x} \in \operatorname{dom}(f)$, then f is convex.

i.e., convex = subgradients everywhere.

Thm 5.4 Hyperplane Separation Theorem

Two nonempty convex sets can be separated by a hyperplane if their (relative) interiors do not intersect.

Thm 5.5 Differentiability of convex functions

A convex function f is differentiable almost everywhere on $\operatorname{dom}(f)$.

Lemma 5.6 Convex and Lipschitz continuity $\Longleftrightarrow$ Bounded Subgradients

Let $f:\operatorname{dom}(f) \to \mathbb{R}$ be convex, $\operatorname{dom}(f)$ open, $B \in \mathbb{R}_+$. Then the following two statements are equivalent:

(i) $\|\mathbf{g}\|_2 \le B$ for all $\mathbf{x} \in \operatorname{dom}(f)$ and all $\mathbf{g} \in \partial f(\mathbf{x})$.

(ii) $|f(\mathbf{x}) - f(\mathbf{y})| \le B\|\mathbf{x} - \mathbf{y}\|_2$ for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom}(f)$.

More generally, let $f: S \to \mathbb{R}$ be a convex function. Then, f is $L-$ Lipschitz over S with respect to a norm $\|\cdot\|$ iff for all $\mathbf{w} \in S$ and $\mathbf{z} \in \partial f(\mathbf{w})$ we have that $\|\mathbf{z}\|_\star \le L$, where $\|\cdot\|_\star$ is the dual norm.

Lemma 5.7 Subgradient optimality condition

Suppose that $f:\operatorname{dom}(f) \to \mathbb{R}$ and $\mathbf{x} \in \operatorname{dom}(f)$. If $0 \in \partial f(\mathbf{x})$, then $\mathbf{x}$ is a global min.

Calculus of Subgradient and Subdifferential

- Conic combination: Let $h(\mathbf{x}) = \beta_1 f_1(\mathbf{x}) + \beta_2 f_2(\mathbf{x})$ with $\beta_1, \beta_2 \ge 0$, then $\partial h(\mathbf{x}) = \beta_1 \partial f_1(\mathbf{x}) + \beta_2 \partial f_2(\mathbf{x})$
- Affine transformation: Let $h(\mathbf{x}) = f(A\mathbf{x} + b)$, then $\partial h(\mathbf{x}) = A^\top \partial f(A\mathbf{x} + b)$
- Pointwise maximum: Let $h(\mathbf{x}) = \max_{i=1,\ldots,m} f_i(\mathbf{x})$, then $\partial h(x) = \operatorname{conv}\{\partial f_i(\mathbf{x}) : i \text{ such that } f_i(\mathbf{x}) = h(\mathbf{x})\}$ (convex hull)

Remark

Negative subgradient may not be a descent direction.

Convex Nonsmooth Problem Setting

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in X \end{array}$$

Assume that

- $f(\mathbf{x})$ is convex and B-Lipschitz continuous on X:
$$|f(\mathbf{x}) - f(\mathbf{y})| \le B\|\mathbf{x} - \mathbf{y}\|_2, \quad \forall \mathbf{x}, \mathbf{y} \in X.$$
This implies that $\|\mathbf{g}\|_2 \le B$ for any $\mathbf{g} \in \partial f(x)$.
- X is convex and compact: $R := \max_{\mathbf{x},\mathbf{y} \in X} \|\mathbf{x} - \mathbf{y}\|_2 < +\infty$.

Denote X^* as the optimal set such that $X^* \neq \emptyset$.
Denote f^* as the optimal value such that $f^* < \infty$.

Subgradient Descent

Choose $\mathbf{x}_1 \in \mathbb{R}^d$.

$$\begin{aligned} \mathbf{x}_{t+1} &:= \Pi_X(\mathbf{x}_t - \gamma_t \mathbf{g}_t) \\ &= \operatorname*{argmin}_{\mathbf{x} \in X} \left\{\frac{1}{2}\|\mathbf{x} - \mathbf{x}_t\|_2^2 + \langle \gamma_t \mathbf{g}_t, \mathbf{x}\rangle\right\}, \mathbf{g}_t \in \partial f(\mathbf{x}_t) \end{aligned}$$

- $\mathbf{g}_t$ is a subgradient of f at $\mathbf{x}_t$.
- $\gamma_t > 0$ is a proper (time-varying) stepsize.
- $\Pi_X(\mathbf{y}) := \operatorname{argmin}_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{y}\|_2^2$ is the projection.
- When f is differentiable and $X = \mathbb{R}^d$, this reduces to Gradient Descent.
- When f is differentiable and $X \subset \mathbb{R}^d$, this reduces to Projected Gradient Descent.
- When f is non-differentiable, we see that it is not always a descent method.

Choices of Stepsize

- Constant stepsize: $\gamma_t = \gamma$
- Scaled stepsize: $\gamma_t = \frac{\gamma}{\|g_t\|_2}$
- Non-summable but diminishing stepsize: $\gamma_t \to 0$ and $\sum_{t=1}^\infty \gamma_t = +\infty$, e.g.: $\gamma = \mathcal{O}\left(\frac{1}{\sqrt{t}}\right)$
- Square summable stepsize: $\sum_{t=1}^\infty \gamma_t^2 < +\infty$ and $\sum_{t=1}^\infty \gamma_t = +\infty$, e.g.: $\gamma = \mathcal{O}\left(\frac{1}{t}\right)$
- Polyak's stepsize: $\gamma_t = \frac{f(\mathbf{x}_t) - f^*}{\|\mathbf{g}_t\|_2^2}$, where f^* is the optimal value.

Lemma 5.8 Basic Descent Lemma

If f is convex (and B-Lipschitz), then for any optimal solution $\mathbf{x}^* \in X^*$,

$$\left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|_2^2 \le \left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2 - 2\gamma_t(f(\mathbf{x}_t) - f^*) + \gamma_t^2\|\mathbf{g}_t\|_2^2$$

$$\Rightarrow \left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|_2^2 \le \left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2 + \gamma_t^2\|\mathbf{g}_t\|_2^2$$

$$\Rightarrow \left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|_2^2 \le \left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2 + \gamma_t^2 B^2$$

$$\Rightarrow \left\|\mathbf{x}_T - \mathbf{x}^*\right\|_2^2 \le \left\|\mathbf{x}_{t_k} - \mathbf{x}^*\right\|_2^2 + B^2 \sum_{t=t_k}^{T-1} \gamma_t^2$$

Thm 5.9 Main Theorem on Convergence

If f is convex, then the subgradient method satisfies:

$$\min_{1 \le t \le T} f(\mathbf{x}_t) - f^* \le \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|_2^2 + \sum_{t=1}^T \gamma_t^2\|\mathbf{g}_t\|_2^2}{2\sum_{t=1}^T \gamma_t}$$

Asymptotic Convergence under Different Stepsizes

Recall $\|\mathbf{x}_t - \mathbf{x}^*\| \le R^2$ and $\|\mathbf{g}_t\|_2^2 \le B^2$

- Constant $\gamma_t \equiv \gamma$: $\liminf_{t\to\infty} f(\mathbf{x}_t) \le f^* + B^2\gamma/2$
- Scaled $\gamma_t = \frac{\gamma}{\|g_t\|_2}$: $\liminf_{t\to\infty} f(\mathbf{x}_t) \le f^* + B\gamma/2$
- Square-summable $\sum_{t=1}^\infty \gamma_t^2 < +\infty$ and $\sum_{t=1}^\infty \gamma_t = +\infty$: $\liminf_{t\to\infty} f(\mathbf{x}_t) = f^*$
- Diminishing $\gamma_t \to 0$ and $\sum_{t=1}^\infty \gamma_t = +\infty$: $\liminf_{t\to\infty} f(\mathbf{x}_t) = f^*$

Subgradient Descent Convergence under Polyak's Stepsizes

Minimizing the surrogate func in Lemma 5.8 yields the optimal stepsize (Polyak): $\gamma_t = \frac{f(\mathbf{x}_t) - f^*}{\|\mathbf{g}_t\|_2^2}$

This guarantees strict error reduction (*):

$$\left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|_2^2 \le \left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2 - \frac{(f(\mathbf{x}_t) - f^*)^2}{\|g(\mathbf{x}_t)\|_2^2}$$

It follows that $f(\mathbf{x}_t) \to f^*$ and $\{\mathbf{x}_t\} \to \mathbf{x}^*$.

Polyak's Stepsize is useful when the optimal value f^* is known, but minimizer $\mathbf{x}^*$ is unknown.

In practice, the opt value is often not available. One can replace f^* by an online estimate, e.g., $\hat{f}_t := \min_{0 \le \tau \le t} f(\mathbf{x}_\tau) - \delta$

- Assume f is convex and B-Lipscitz, then (*) implies
$$\min_{1 \le t \le T} f(\mathbf{x}_t) - f(\mathbf{x}^*) \le \frac{B\|\mathbf{x}_1 - \mathbf{x}^*\|_2}{\sqrt{T}}$$
(Proof see Hw5 Ex4.1)
- Asuume f is μ strongly convex and B-Lipscitz.
  - In the case where f is non-differentiable, the definition of strong convexity we saw in lecture no longer applies. However, we can still define strong convexity in this setting as follows: For $\mu > 0$, A function f is said to be μ-strongly convex if the function $f_\mu(x) := f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|^2$ is convex. Let f be μ-strongly convex, then for all $\mathbf{x}, \mathbf{y}$ in the domain and for all $\mathbf{g} \in \partial f(\mathbf{x})$ we have $f(\mathbf{y}) \ge f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2$.

  Then (*) implies
$$\min_{1 \le t \le T} f(\mathbf{x}_t) - f(\mathbf{x}^*) \le \frac{4B^2}{\mu T}$$
(Proof see Hw5 Ex4.2)

Corollary 5.10 Convergence Rate for Convex Lipschitz Problem

If f is convex and B-Lipschitz continuous, and X is convex compact with diameter R. Let $\gamma_t \equiv \frac{R}{B\sqrt{T}}$ or $\gamma_t = \frac{R}{B\sqrt{t}}$, then

$$\min_{1\le t\le T} f(\mathbf{x}_t) - f^* \le \mathcal{O}\left(\frac{BR}{\sqrt{T}}\right)$$

- Subgradient method **converges sublinearly**.
- For an accuracy $\epsilon > 0$, need $\mathcal{O}\left(\frac{B^2R^2}{\epsilon^2}\right)$ number of iterations or subgradients.

Strongly Convex and Lipschitz Problem

We now consider an even nicer problem class: $f(\mathbf{x})$ is μ-**strongly convex** on X with $\mu > 0$:

$$f(\mathbf{x}) \ge f(\mathbf{y}) + \nabla f(\mathbf{y})^\top(\mathbf{x}-\mathbf{y}) + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|_2^2. \quad \forall \mathbf{x},\mathbf{y}\in X$$

Lemma 5.11 Descent Lemma

$$\left\|\mathbf{x}_{t+1}-\mathbf{x}^*\right\|_2^2 \le (1-\mu\gamma_t)\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2 - 2\gamma_t(f(\mathbf{x}_t)-f^*) + \gamma_t^2\|\mathbf{g}_t\|_2^2$$

Thm 5.12

Let f be μ-**strongly convex and B-Lipschitz continuous** on X, then with $\gamma_t = \frac{2}{\mu(t+1)}$, we have

$$\min_{1\le t\le T} f(\mathbf{x}_t) - f^* \le \frac{2B^2}{\mu\cdot(T+1)}$$

Summary of Subgradient Method

	Cvx	Strongly Cvx
Convergence rate	$O\left(\frac{B\cdot R}{\sqrt{t}}\right)$	$O\left(\frac{B^2}{\mu t}\right)$
Subgrad complexity	$O\left(\frac{B\cdot R}{\epsilon^2}\right)$	$O\left(\frac{B^2}{\mu\epsilon}\right)$

$$B := \sup_{\mathbf{x}\in X}\frac{|f(\mathbf{x})-f(\mathbf{y})|}{\|\mathbf{x}-\mathbf{y}\|_2}, R := \max_{\mathbf{x},\mathbf{y}\in X}\|\mathbf{x}-\mathbf{y}\|_2$$

Subgradient Descent vs. Gradient Descent

Setting	Algo	Cvx	Strongly Cvx
Nonsmooth	Subgrad	$O\left(\frac{B\cdot R}{\sqrt{t}}\right)$	$O\left(\frac{B^2}{\mu t}\right)$
Smooth	GD	$O\left(\frac{L\cdot R^2}{t}\right)$	$O\left(\left(1-\frac{\mu}{L}\right)^t\right)$
	AGD	$O\left(\frac{L\cdot R^2}{t^2}\right)$	$O\left(\left(1-\sqrt{\frac{\mu}{L}}\right)^t\right)$

Lower Complexity Bound for Nonsmooth Cvx Opt

In the worst case, the sublinear rates $O(1/\sqrt{t})$ and $O(1/t)$ for convex and strongly convex Lipschitz problems **cannot be improved**, for algorithms using only subgradient oracles. Subgradient descent is "**optimal**" for such problem classes.

Thm 5.13 (Nemirovski & Yudin, 1983)

For any $1\le t\le d, \mathbf{x}_1\in\mathbb{R}^d$, there exists a B-Lipschitz continuous and convex function f, a convex set X with diameter R, such that for any first-order method that generates:

$$\mathbf{x}_t \in \mathbf{x}_1 + \operatorname{span}(\mathbf{g}_1,\ldots,\mathbf{g}_{t-1}), \mathbf{g}_i\in\partial f(\mathbf{x}_i), i=1,\ldots,t-1$$

We have $\min_{1\le s\le t} f(\mathbf{x}_s) - f^* \ge \frac{B\cdot R}{4(1+\sqrt{t})}$

6. Stochastic Optimization

General Stochastic Optimization (SO) Problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) := \mathbb{E}_\xi[f(\mathbf{x},\xi)]$$

- ξ is a random vector with support $\Xi\subset\mathbb{R}^m$ and distribution P.
- For simplicity, assume $f(\mathbf{x},\xi)$ is continuously differentiable for any $\xi\in\Xi$.

Finite Sum Problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})$$

$F(\mathbf{x}) = \mathbb{E}_\xi\left[f_\xi(\mathbf{x})\right]$, where ξ is uniformly distributed over $\{1,2,\ldots,n\}$.

SO Pros & Cons

Pros: (1) Faster, (2) Memory efficient, (3) Avoid overfitting, help generalization.
Cons: (1) Lack of guarantee (2) Stuck at local solution, (3) Require strong assumptions

SGD for Finite Sum Problem

Sample $i_t\in[n]$ **uniformly** at random

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t\nabla f_{i_t}(\mathbf{x}_t)$$

Unbiasedness: $\mathbb{E}_{i_t}\left[\nabla f_{i_t}(\mathbf{x})\right] = \frac{1}{n}\sum_{i=1}^n\nabla f_i(\mathbf{x}) = \nabla F(\mathbf{x})$
Each iteration is $\mathcal{O}(n)$ cheaper than full GD.

Vanilla Analysis of Finite Sum Problem

If i is chosen at step t, then we have

$$\nabla f_i(\mathbf{x}_t)^\top(\mathbf{x}_t-\mathbf{x}^*) = \frac{\gamma_t}{2}\|\nabla f_i(\mathbf{x}_t)\|_2^2 + \frac{1}{2\gamma_t}\left(\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2 - \left\|\mathbf{x}_{t+1}-\mathbf{x}^*\right\|_2^2\right)$$

SGD for General Stochastic Optimization

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t\nabla f(\mathbf{x}_t,\xi_t), \text{ where } \xi_t \overset{iid}{\sim} P(\xi)$$

Boundedness of Stochastic Gradients

Let $F(\mathbf{x}) = \mathbb{E}[f(\mathbf{x},\xi)]$, where $f(\mathbf{x},\xi)$ is convex and L-smooth for any realization of ξ. Define $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}} F(\mathbf{x})$. Then we have

$$\mathbb{E}\left[\left\|\nabla f(\mathbf{x},\xi) - \nabla f(\mathbf{x}^*,\xi)\right\|_2^2\right] \le 2L[F(\mathbf{x})-F(\mathbf{x}^*)]$$

$$\mathbb{E}\left[\|\nabla f(\mathbf{x},\xi)\|_2^2\right] \le 4L[F(\mathbf{x})-F(\mathbf{x}^*)] + 2\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^*,\xi)\right\|_2^2\right]$$

(**Proof see Hw6 Ex1&2**)
Unbiasedness: $\mathbb{E}[\nabla f(\mathbf{x}_t,\xi_t)\,|\,\mathbf{x}_t] = \nabla F(\mathbf{x}_t)$ under mild regularity conditions
We always assume stochastic gradient is **unbiased**.
Note SGD is **not** a monotonic descent method.

Stepsize (or Learning Rate)

- If use fixed stepsize for SGD as in GD, SGD will not converge to the optimal solution (almost surely).
- Stepsize should decrease to $0, \gamma_t\to 0$
- For example, use polynomial rate $\gamma_t = \mathcal{O}(t^{-a})$ with some $a>0$
- In practice, use the form $\gamma_t = \frac{\gamma_0}{1+\beta t}$ and tune hyperparameters γ_0,β
- In deep learning, often adopt **step decay** - drop the learning rate by a factor every few epochs.

Simple Improvements of SGD

Mini-batch SGD

Use b random samples to construct gradient estimator

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t\cdot\frac{1}{b}\sum_{j\in J,|J|=b}\nabla f\left(\mathbf{x}_t,\xi_j\right)$$

SGD with iterate averaging

$$\bar{\mathbf{x}}_t = \frac{1}{t}\sum_{\tau=1}^t\mathbf{x}_\tau$$

Averaging and mini-batch sampling can help reduce the variance.
Still, SGD can be **very sensitive to the choice of stepsize**.

Limitations of SGD

- learning rate tuning
- uniform learning rate for all coordinates

Convergence for Stochastic Convex Problems

Thm 6.1 (Convex, weighted averaging)

Suppose $F(\mathbf{x})$ is **convex** and $\mathbb{E}\left[\|\nabla f(\mathbf{x},\xi)\|_2^2\right]\le B^2,\forall\mathbf{x}$.
Then SGD satisfies that

$$\mathbb{E}[F(\hat{\mathbf{x}}_T)-F(\mathbf{x}^*)] \le \frac{R^2+B^2\sum_{t=1}^T\gamma_t^2}{2\sum_{t=1}^T\gamma_t}$$

where $\hat{\mathbf{x}}_T := \sum_{t=1}^T\gamma_t\mathbf{x}_t/\sum_{t=1}^T\gamma_t$ and $\|\mathbf{x}_1-\mathbf{x}^*\|_2\le R$.

- If $\gamma_t\equiv\frac{R}{B\sqrt{T}}, \mathbb{E}[F(\hat{\mathbf{x}}_T)-F(\mathbf{x}^*)] = \mathcal{O}\left(\frac{BR}{\sqrt{T}}\right)$.
- This further implies the $\mathcal{O}\left(1/\epsilon^2\right)$ sample complexity required by SGD.

Thm 6.2 (Strong convex, diminishing stepsize, last iterate)

Assume $F(\mathbf{x})$ is μ-**strongly convex** and $\mathbb{E}\left[\|\nabla f(\mathbf{x},\xi)\|_2^2\right]\le B^2,\forall\mathbf{x}$, then SGD with $\gamma_t = \frac{\gamma}{t}\left(\gamma>\frac{1}{2\mu}\right)$ satisfies

$$\mathbb{E}\left[\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2\right] \le \frac{C(\gamma)}{t}$$

where $C(\gamma) = \max\left\{\frac{\gamma^2B^2}{2\mu\gamma-1},\|\mathbf{x}_1-\mathbf{x}^*\|_2^2\right\}$

- If F is also L-smooth, this further implies that $\mathbb{E}[F(\mathbf{x}_t)-f(\mathbf{x}^*)] = \mathcal{O}\left(\frac{L\cdot C(\gamma)}{t}\right)$.
- The sample complexity required by SGD is $\mathcal{O}(1/\epsilon)$ in this case.

SGD under Constant Stepsize (Thm 5.3)

Assume that (1) $F(\mathbf{x}) := \mathbb{E}[f(\mathbf{x},\xi)]$ is μ-strongly convex and L-smooth; (2) The unbiased estimator satisfies that for all $\mathbf{x}$:

$$\mathbb{E}\left[\|\nabla f(\mathbf{x},\xi)\|_2^2\right] \le \sigma^2 + c\|\nabla F(\mathbf{x})\|_2^2$$

Under the above assumption, SGD with $\gamma_t=\gamma\le\frac{1}{Lc}$ achieves:

$$\mathbb{E}[F(\mathbf{x}_t)-F(\mathbf{x}^*)] \le \frac{\gamma L\sigma^2}{2\mu} + (1-\mu\gamma)^{t-1}(F(\mathbf{x}_1)-F(\mathbf{x}^*))$$

- With constant stepsize, SGD converges **linearly to a neighborhood** around $\mathbf{x}^*$.
- **Accuracy-convergence trade-off**: Smaller stepsize γ implies better solution but slower rate.
- **Strong Growth Condition**: when $\sigma^2=0$, i.e., $\mathbb{E}\left[\|\nabla f(\mathbf{x},\xi)\|_2^2\right]\le c\|\nabla F(\mathbf{x})\|_2^2$, SGD with constant stepsize **converges to the global optimum at a linear rate**.
  - Consider $F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})$, strong growth condition implies interpolation: at optimal solution $\mathbf{x}^*, \nabla f_i(\mathbf{x}^*)=0,\forall i$.
  - Strong growth condition holds when F is smooth and satisfies PL inequality.
  - Examples: linear regression or overparametrized neural network in the realizable case.

SGD Summary

	Cvx	Strongly Cvx
Convergence rate	$O\left(\frac{1}{\sqrt{t}}\right)$	$O\left(\frac{1}{t}\right)$
Sample complexity	$O\left(\frac{1}{\epsilon^2}\right)$	$O\left(\frac{1}{\epsilon}\right)$

Lower Complexity Bound for Stochastic Optimization

In the worst case, the sample complexity $\mathcal{O}(1/\epsilon^2)$ and $\mathcal{O}(1/\epsilon)$ for convex and strongly convex Lipschitz problems **cannot be improved**, for algorithms using only stochastic oracles.
Stochastic Oracle: given input $\mathbf{x}$, stochastic oracle returns $G(\mathbf{x},\xi)$ such that

$$\mathbb{E}[G(\mathbf{x},\xi)] \in \partial f(\mathbf{x}) \text{ and } \mathbb{E}\left[\|G(\mathbf{x},\xi)\|_p^2\right] \le M^2$$

for some positive constant M and some $p \in [1,\infty]$.
SGD is **optimal** for such problem classes.

Popular Variants of SGD

Momentum SGD

$$\begin{cases} \mathbf{m}_t = \alpha \mathbf{m}_{t-1} + (1-\alpha)\nabla f(\mathbf{x}_t,\xi_t) \\ \mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t \mathbf{m}_t \end{cases}$$

AdaGrad

$$\begin{cases} \mathbf{v}_t = \mathbf{v}_{t-1} + \nabla f(\mathbf{x}_t,\xi_t)^{\odot 2} \\ \mathbf{x}_{t+1} = \mathbf{x}_t - \frac{\gamma_0}{\epsilon+\sqrt{\mathbf{v}_t}} \odot \nabla f(\mathbf{x}_t,\xi_t) \end{cases}$$

RMSProp

$$\begin{cases} \mathbf{v}_t = \beta \mathbf{v}_{t-1} + (1-\beta)\nabla f(\mathbf{x}_t,\xi_t)^{\odot 2} \\ \mathbf{x}_{t+1} = \mathbf{x}_t - \frac{\gamma_0}{\varepsilon+\sqrt{\mathbf{v}_t}} \odot \nabla f(\mathbf{x}_t,\xi_t) \end{cases}$$

ADAM

ADAM $\approx$ RMSProp + Momentum

$$\begin{cases} \mathbf{v}_t = \beta \mathbf{v}_{t-1} + (1-\beta)\nabla f(\mathbf{x}_t,\xi_t)^{\odot 2} \\ \mathbf{m}_t = \alpha \mathbf{m}_{t-1} + (1-\alpha)\nabla f(\mathbf{x}_t,\xi_t) \\ \mathbf{x}_{t+1} = \mathbf{x}_t - \frac{\gamma_0}{\varepsilon+\sqrt{\tilde{\mathbf{v}}_t}} \odot \tilde{\mathbf{m}}_t \end{cases}$$

- Exponential decay of previous information $\mathbf{m}_t, \mathbf{v}_t$.
- Note $\tilde{\mathbf{v}}_t = \frac{\mathbf{v}_t}{1-\beta^t}$ and $\tilde{\mathbf{m}}_t = \frac{\mathbf{m}_t}{1-\alpha^t}$ are bias-corrected estimates.
- In practice, α and β are chosen to be close to 1 .

Generic Adaptive Scheme

$$\begin{aligned} \mathbf{g}_t &= \nabla f(\mathbf{x}_t,\xi_t) \\ \mathbf{m}_t &= \phi_t(\mathbf{g}_1,\dots,\mathbf{g}_t) \\ \mathbf{V}_t &= \psi_t(\mathbf{g}_1,\dots,\mathbf{g}_t) \\ \hat{\mathbf{x}}_t &= \mathbf{x}_t - \alpha_t \mathbf{V}_t^{-1/2}\mathbf{m}_t \\ \mathbf{x}_{t+1} &= \underset{\mathbf{x}\in X}{\operatorname{argmin}} \left\{(\mathbf{x}-\hat{\mathbf{x}}_t)^T \mathbf{V}_t^{1/2}(\mathbf{x}-\hat{\mathbf{x}}_t)\right\} \end{aligned}$$

- SGD: $\phi_t(\mathbf{g}_1,\dots,\mathbf{g}_t) = \mathbf{g}_t$, $\psi_t(\mathbf{g}_1,\dots,\mathbf{g}_t) = \mathbf{I}$
- AdaGrad: $\phi_t(\mathbf{g}_1,\dots,\mathbf{g}_t) = \mathbf{g}_t$, $\psi_t(\mathbf{g}_1,\dots,\mathbf{g}_t) = \frac{\operatorname{diag}(\sum_{\tau=1}^t \mathbf{g}_\tau^2)}{t}$
- Adam: $\phi_t(\mathbf{g}_1,\dots,\mathbf{g}_t) = (1-\beta_1)\sum_{\tau=1}^t \beta_1^{t-\tau}\mathbf{g}_\tau$, $\psi_t(\mathbf{g}_1,\dots,\mathbf{g}_t) = (1-\beta_2)\operatorname{diag}\left(\sum_{\tau=1}^t \beta_2^{t-\tau}\mathbf{g}_\tau^2\right)$. In other words, $\mathbf{m}_t = \beta_1\mathbf{m}_{t-1} + (1-\beta_1)\mathbf{g}_t$, $\mathbf{V}_t = \beta_2\mathbf{V}_{t-1} + (1-\beta_2)\operatorname{diag}(\mathbf{g}_t^2)$.

7. Variance-reduced Stochastic Methods

The Non-Convergence of Adam

Counterexample: consider a one-dim problem:

$$X = [-1,1],\ f(x,\xi) = \begin{cases} Cx, & \text{if } \xi = 1 \\ -x, & \text{if } \xi = 0 \end{cases}$$

$$\mathbb{P}(\xi = 1) = p = \frac{1+\delta}{C+1}$$

- Here $F(x) = \mathbb{E}[f(x,\xi)] = \delta x$ and $x^* = -1$.
- Adam step is $x_{t+1} = x_t - \gamma_0\Delta_t$ with $\Delta_t = \frac{\alpha m_t + (1-\alpha)g_t}{\sqrt{\beta v_t + (1-\beta)g_t^2}}$
- For large enough $C > 0$, one can show $\mathbb{E}[\Delta_t] \le 0$.
- Adam steps keep drifting away from the optimal solution $x^* = -1$.

SGD vs. GD for Finite Sum Problem

Complexity for smooth and strongly-convex problems: $\kappa := L/\mu$.

	iter complexity	per-iter cost	total cost
GD	$\mathcal{O}(\kappa \cdot \ln\frac{1}{\epsilon})$	$\mathcal{O}(n)$	$\mathcal{O}(n\kappa\ln\frac{1}{\epsilon})$
SGD	$\mathcal{O}(\frac{\kappa}{\epsilon})$	$\mathcal{O}(1)$	$\mathcal{O}(\frac{\kappa}{\epsilon})$

- GD converges faster but with expensive iter cost.
- SGD converges slowly but with cheap iter cost.
- SGD is more appealing for large n and moderate accuracy ϵ.

SGD vs. GD vs. VR Methods

Algo	# of Iterations	Per-iteration Cost
GD	$O\left(\kappa\log\frac{1}{\epsilon}\right)$	$O(n)$
SGD	$O\left(\frac{\kappa}{\epsilon}\right)$	$O(1)$
VR	$O\left((n+\kappa)\log\frac{1}{\epsilon}\right)$	$O(1)$

Classical Variance Reduction Techniques

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})$$

Mini-batching: Use the average of gradients from a random subset

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t\frac{1}{|B_t|}\sum_{i\in B_t}\nabla f_i(\mathbf{x}_t)$$

Note: VR comes at a computational cost.
Momentum: add momentum to the gradient step

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t\hat{\mathbf{m}}_t,\ \text{where } \hat{\mathbf{m}}_t = c\cdot\sum_{\tau=1}^t \alpha^{t-\tau}\nabla f_{i_\tau}(\mathbf{x}_\tau)$$

Note: Here $\mathbf{m}_t$ is the weighted average of the past stochastic gradients.

Modern Variance Reduction Technique

Suppose X is positively correlated with Y and we can compute $\mathbb{E}[Y]$.
Point Estimator:

$$\begin{aligned} \hat{\Theta}_\alpha &= \alpha(X-Y) + \mathbb{E}[Y], \quad (0\le\alpha\le 1) \\ \mathbb{E}\left[\hat{\Theta}_\alpha\right] &= \alpha\mathbb{E}[X] + (1-\alpha)\mathbb{E}[Y] \\ \mathbb{V}\left[\hat{\Theta}_\alpha\right] &= \alpha^2(\mathbb{V}[X] + \mathbb{V}[Y] - 2\operatorname{Cov}[X,Y]) \end{aligned}$$

If cov is sufficiently large, then $\mathbb{V}\left[\hat{\Theta}_\alpha\right] \le \mathbb{V}[X]$.

Variance Reduction Techniques for Finite Sum Problems

Goal: estimate $\theta = \nabla F(\mathbf{x}_t), X = \nabla f_{i_t}(\mathbf{x}_t)$

- SGD: $\mathbf{g}_t = \nabla f_{i_t}(\mathbf{x}_t)$ $[\alpha = 1, Y = 0]$
- SAG: $\mathbf{g}_t = \frac{1}{n}\left(\nabla f_{i_t}(\mathbf{x}_t) - \mathbf{v}_{i_t}\right) + \frac{1}{n}\sum_{i=1}^n \mathbf{v}_i$ $\left[\alpha = \frac{1}{n}, Y = \mathbf{v}_{i_t}\right]$
- SAGA: $\mathbf{g}_t = \left(\nabla f_{i_t}(\mathbf{x}_t) - \mathbf{v}_{i_t}\right) + \frac{1}{n}\sum_{i=1}^n \mathbf{v}_i$ $\left[\alpha = 1, Y = \mathbf{v}_{i_t}\right]$
  Here $\{\mathbf{v}_i, i = 1,\dots,n\}$ are the past stored gradients for each component.
- SVRG: $\mathbf{g}_t = \nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\tilde{\mathbf{x}}) + \nabla F(\tilde{\mathbf{x}})$ $\left[\alpha = 1, Y = \nabla f_{i_t}(\tilde{\mathbf{x}})\right]$

Stochastic Average Gradient (SAG)

Idea: keep track of the average of $\mathbf{v}_i$ as an estimate of the full gradient

$$\mathbf{g}_t = \frac{1}{n}\sum_{i=1}^n \mathbf{v}_i^t \quad\approx\quad \frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_t) = \nabla F(\mathbf{x}_t)$$

The past gradients are updated as:

$$\mathbf{v}_i^t = \begin{cases} \nabla f_{i_t}(\mathbf{x}_t), & \text{if } i = i_t \\ \mathbf{v}_i^{t-1}, & \text{if } i \ne i_t \end{cases}$$

Equivalently, we have

$$\mathbf{g}_t = \mathbf{g}_{t-1} - \frac{1}{n}\underbrace{\mathbf{v}_{i_t}^{t-1}}_{Y} + \frac{1}{n}\underbrace{\nabla f_{i_t}(\mathbf{x}_t)}_{X}$$

$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{\gamma}{n}\sum_{i=1}^n \mathbf{v}_i^t$,
where $\mathbf{v}_i^t = \begin{cases} \nabla f_{i_t}(\mathbf{x}_t), & \text{if } i = i_t \\ \mathbf{v}_i^{t-1}, & \text{otherwise} \end{cases}$
Biased gradient; Cheap iteration cost; $\mathcal{O}(nd)$ memory cost; Hard to analyze.

- Linear convergence: The first stochastic methods to enjoy linear rate using a constant stepsize for strongly-convex and smooth objectives.
- Memory cost: $O(n)$ times higher than SGD/SVRG
- Per-iteration cost: one gradient evaluation
- Total complexity: $O\left((n+\kappa_{\max})\log\left(\frac{1}{\epsilon}\right)\right)$

SAGA

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma\left[\left(\nabla f_{i_t}(\mathbf{x}_t) - \mathbf{v}_{i_t}^{t-1}\right) + \frac{1}{n}\sum_{i=1}^n \mathbf{v}_i^{t-1}\right]$$

- **Unbiased** update, while **SAG is biased**
- Same $\mathcal{O}(nd)$ memory cost as SAG
- Similar linear convergence rate as SAG

Stochastic Variance Reduced Gradient (SVRG)

Intuition: the closer $\tilde{\mathbf{x}}$ is to $\mathbf{x}_t$, the smaller the variance of the gradient estimator

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{g}_t - \nabla F(\mathbf{x}_t)\|^2\right] &\le \mathbb{E}\left[\left\|\nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\tilde{\mathbf{x}})\right\|^2\right] \\ &\le L_{\max}^2\|\mathbf{x}_t - \tilde{\mathbf{x}}\|^2 \end{aligned}$$

Two-loop structure:

- Outer loop: update reference point and compute its full gradient at $O(n)$ cost
- Inner loop: update iterates with variance-reduced gradient for m steps
- Total of $O(n+2m)$ component gradient evaluations at each epoch

Compare to SAG/SAGA
Pro: Cheap memory cost, no need to store past gradients or past iterates
Con: More parameter tuning, two gradient computation per iteration

Thm 7.1 Convergence of SVRG

Assume each $f_i(\mathbf{x})$ is convex and L_i-smooth, $F(\mathbf{x})$ is μ-strongly convex. Assume m is sufficiently large and $\eta < \frac{1}{2L_{\max}}$ such that $\rho = \frac{1}{\mu\eta(1-2\eta L_{\max})m} + \frac{2\eta L_{\max}}{1-2\eta L_{\max}} < 1$, then

$$\mathbb{E}[F(\tilde{\mathbf{x}}^s) - F(\mathbf{x}^*)] \le \rho^s\left[F\left(\tilde{\mathbf{x}}^0\right) - F(\mathbf{x}^*)\right]$$

Linear convergence: choose $m = \mathcal{O}\left(\frac{L_{\max}}{\mu}\right), \eta = \mathcal{O}\left(\frac{1}{L_{\max}}\right)$ such that $\rho\in\left(0,\frac{1}{2}\right)$.
Total complexity:

$$\mathcal{O}\left((2m+n)\log\frac{1}{\epsilon}\right) = \mathcal{O}\left(\left(n + \frac{L_{\max}}{\mu}\right)\log\frac{1}{\epsilon}\right)$$

Lemma 7.2 Property of Smoothness

Let $F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})$, where each $f_i:\mathbb{R}^d\to\mathbb{R}$ is a convex and L_i-smooth function and F has a global minimum $\mathbf{x}^\star$. Let $L_{\max} = \max\{L_1,\dots,L_n\}$. Then, for any $\mathbf{x}\in\mathbb{R}^d$

$$\frac{1}{n}\sum_{i=1}^n\left\|\nabla f_i(\mathbf{x}) - \nabla f_i\left(\mathbf{x}^\star\right)\right\|_2^2 \le 2L_{\max}\left[F(\mathbf{x}) - F\left(\mathbf{x}^\star\right)\right]$$

Lemma 7.3 Bound of Variance

$\tilde{\mathbf{x}}, \mathbf{x_t} \in \mathbb{R}^d$. Denote $\mathbf{g_t} = \nabla f_{i_t}(\mathbf{x_t}) - \nabla f_{i_t}(\tilde{\mathbf{x}}) + \nabla F(\tilde{\mathbf{x}})$, where i_t is sampled uniformly from $\{1, \dots, n\}$. Then

$$\mathbb{E}\left[\|\mathbf{g}_t\|_2^2\right] \le 4L_{\max}\left[F(\mathbf{x}_t) - F(\mathbf{x}^*) + F(\tilde{\mathbf{x}}) - F(\mathbf{x}^*)\right]$$

8. Nonconvex Functions

Concave functions

f is called **concave** if $-f$ is convex.
For all $\mathbf{x}$, the graph of a differentiable concave function is below the tangent hyperplane at $\mathbf{x}$.
$\Rightarrow$ concave functions are smooth with $L = 0$... but boring from an optimization point of view (no global minimum), gradient descent runs off to infinity

Lemma 8.1 Bounded Hessians ⇒ smooth

Let $f : \mathrm{dom}(f) \to \mathbb{R}$ be twice differentiable, with $X \subseteq \mathrm{dom}(f)$ a convex set, and $\left\|\nabla^2 f(\mathbf{x})\right\| \le L$ for all $\mathbf{x} \in X$, where $\|\cdot\|$ is spectral norm. Then f is smooth with parameter L over X.
Examples:

- all quadratic functions $f(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$
- $f(x) = \sin(x)$ (many global minima)

Thm 8.2 Gradient descent on smooth (not necessarily convex) functions

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable with a global minimum $\mathbf{x}^\star$; furthermore, suppose that f is smooth with parameter L according to Definition 3.2. Choosing stepsize $\gamma := \frac{1}{L}$. Gradient descent yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 \le \frac{2L}{T}\left(f(\mathbf{x}_0) - f\left(\mathbf{x}^\star\right)\right), \quad T > 0$$

.
In particular, $\|\nabla f(\mathbf{x}_t)\|^2 \le \frac{2L}{T}\left(f(\mathbf{x}_0) - f\left(\mathbf{x}^\star\right)\right)$ for some $t \in \{0, \dots, T-1\}$.

Corollary of Thm 8.2

Thm 8.2 implies that

$$\lim_{t\to\infty}\|\nabla f(\mathbf{x}_t)\|^2 = 0.$$

Lemma 8.3 No overshooting

In the smooth setting, and with stepsize $1/L$, gradient descent cannot **overshoot**, i.e. pass a critical point.

Local Optimality Problem

Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^n$ to $\mathbb{R}$.
The problem $\mathrm{LocMin}(\mathcal{F})$ is to decide whether $\mathbf{0}$ is a local minimum of a given function $\phi \in \mathcal{F}$.

Thm (Murty and Kabadi [MK87])

The problem $\mathrm{LOCMIN}(\mathcal{F})$ is **coNP-complete** for the class $\mathcal{F} := \{\phi_{\mathbf{M}} : \mathbf{M} \text{ symmetric}\}$, where the function ϕ_M is defined by

$$\phi_M(\mathbf{x}) = \left(\mathbf{x}^2\right)^\top \mathbf{M}\left(\mathbf{x}^2\right)$$

with $\mathbf{x}^2 = \left(x_1^2, x_2^2, \dots, x_n^2\right)$

Proof outline:
$\mathbf{0}$ is a local minimum **iff** the matrix $\mathbf{M}$ is **copositive**.
Deciding whether $\mathbf{M}$ is copositive is coNP-complete.

Copositive matrices

Lemma

$\mathbf{0}$ is a local minimum of $\left(\mathbf{x}^2\right)^\top \mathbf{M}\left(\mathbf{x}^2\right)$ **iff** $\mathbf{x}^\top \mathbf{M}\mathbf{x} \ge 0$ for all $\mathbf{x} \ge \mathbf{0}$.
A matrix $\mathbf{M}$ satisfying $\mathbf{x}^\top \mathbf{M}\mathbf{x} \ge 0$ for all $\mathbf{x} \ge \mathbf{0}$ is called **copositive**.
If $\mathbf{M}$ is positive semidefinite $\left(\mathbf{x}^\top M\mathbf{x} \ge 0 \text{ for all } \mathbf{x}\right)$, then $\mathbf{M}$ is copositive. The converse is false.

Proof

$\mathbf{0}$ is a local minimum

$\Leftrightarrow \left(\mathbf{x}^2\right)^\top \mathbf{M}\left(\mathbf{x}^2\right) \ge 0$ for all $\mathbf{x}$ in some neighborhood of $\mathbf{0}$

$\Leftrightarrow \mathbf{x}^\top \mathbf{M}\mathbf{x} \ge 0$ for all $\mathbf{x} \ge \mathbf{0}$ in some neighborhood of $\mathbf{0}$

$\Leftrightarrow \mathbf{x}^\top \mathbf{M}\mathbf{x} \ge 0$ for all $\mathbf{x} \ge \mathbf{0}$

Def 8.4 c-balanced

Let $\mathbf{x} > \mathbf{0}$ (componentwise), and let $c \ge 1$ be a real number. $\mathbf{x}$ is called c-**balanced** if $x_i \le c x_j$ for all $1 \le i, j \le d$.
Any initial iterate $\mathbf{x}_0 > \mathbf{0}$ is c-balanced for some (possibly large) c.

Lemma 8.5 Balanced iterates

Let $\mathbf{x} > \mathbf{0}$ be c-balanced with $\prod_k x_k \le 1$. Then for any stepsize $\gamma > 0$, $\mathbf{x}' := \mathbf{x} - \gamma\nabla f(\mathbf{x})$ satisfies $\mathbf{x}' \ge \mathbf{x}$ (componentwise) and is also c-balanced.

Lemma 8.6

Suppose that $\mathbf{x} > \mathbf{0}$ is c-balanced. Then for any $I \subseteq \{1, \dots, d\}$, we have

$$\left(\frac{1}{c}\right)^{|I|}\left(\prod_k x_k\right)^{1-|I|/d} \le \prod_{k\notin I} x_k \le c^{|I|}\left(\prod_k x_k\right)^{1-|I|/d}$$

Lemma 8.7

Let $\mathbf{x} > \mathbf{0}$ be c-balanced with $\prod_k x_k \le 1$. Then

$$\left\|\nabla^2 f(\mathbf{x})\right\|_2 \le \left\|\nabla^2 f(\mathbf{x})\right\|_F \le 3dc^2$$

where $\|\cdot\|_F$ is the Frobenius norm and $\|\cdot\|_2$ the spectral norm.

Lemma 8.8

Let $\mathbf{x} > \mathbf{0}$ be c-balanced with $\prod_k x_k < 1, L = 3dc^2$. Let $\gamma := 1/L$. We already know from Lemma 8.5 that $\mathbf{x}' := \mathbf{x} - \gamma\nabla f(\mathbf{x}) \ge \mathbf{x}$ is c-balanced.
Furthermore, f **is smooth with parameter L over the line segment connecting $\mathbf{x}$ and $\mathbf{x}'$**. Lemma 8.3 (no overshooting) then also yields $\prod_k x_k' \le 1$.

Thm 8.9 Convergence of Balanced Iterates

Let $c \ge 1$ and $\delta > 0$ such that $\mathbf{x}_0 > \mathbf{0}$ is c-balanced with $\delta \le \prod_k(\mathbf{x}_0)_k < 1$. Choosing stepsize $\gamma = \frac{1}{3dc^2}$, gradient descent satisfies

$$f(\mathbf{x}_T) \le \left(1 - \frac{\delta^2}{3c^4}\right)^T f(\mathbf{x}_0), \quad T \ge 0$$

Error converges to 0 exponentially fast.

Corollary of Thm 8.9

The sequence $(\mathbf{x}_T)_{T\ge 0}$ of iterates in Thm 8.9 converges to a an optimal solution $\mathbf{x}^\star$.

9. The Frank-Wolfe Algorithm

Linear minimization oracle

Given $\mathbf{g} \in \mathbb{R}^d$,

$$\mathrm{LMO}_X(\mathbf{g}) := \underset{\mathbf{z}\in X}{\operatorname{argmin}}\ \mathbf{g}^\top \mathbf{z}$$

is any minimizer of the linear function $\mathbf{g}^\top \mathbf{z}$ over X. We assume that a minimizer exists whenever we apply the oracle. If X is closed and bounded, this is guaranteed.

Frank-Wolfe Algorithm

Given an initial feasible point $\mathbf{x}_0 \in X$, and (time-dependent) stepsizes $\gamma_t \in [0, 1]$, repeat the following for $t = 0, 1, \dots$:

$$\mathbf{s} := \mathrm{LMO}_X(\nabla f(\mathbf{x}_t))$$
$$\mathbf{x}_{t+1} := (1 - \gamma_t)\mathbf{x}_t + \gamma_t \mathbf{s}$$

Let $\mathbf{y} := \mathbf{x}_{t+1} = \mathbf{x} + \gamma(\mathbf{s} - \mathbf{x})$, then

$$\frac{1}{\gamma^2}\|\mathbf{y} - \mathbf{x}\|^2 = \|\mathbf{s} - \mathbf{x}\|^2$$

Atoms

The Frank-Wolfe algorithm is particularly useful when X is the convex hull of a finite or otherwise nice set of points $\mathcal{A}$ (the **atoms**, or **extreme points**), $X = \mathrm{conv}(\mathcal{A})$.

- $\mathrm{LMO}_X(\mathbf{g}) = \operatorname{argmin}_{\mathbf{z}\in X} \mathbf{g}^\top \mathbf{z}$ is always attained by some atom.
- This may significantly simplify the search for $\mathbf{s} = \mathrm{LMO}_X(\mathbf{g})$.

Example: Spectahedron

Hazan's algorithm: an application of the Frank-Wolfe algorithm to semidefinite programming. $\mathrm{LMO}_X(\mathbf{G})$:

$$\begin{array}{ll} \text{argmin} & \mathbf{G} \bullet \mathbf{Z} \\ \text{subject to} & \mathrm{Tr}(\mathbf{Z}) = 1 \\ & \mathbf{Z} \succeq 0 \end{array}$$

- X is the spectahedron, the set of all (symmetric) positive semidefinite matrices $\mathbf{Z} \in \mathbb{R}^{d\times d}$ of trace 1. **Spectahedron**: $X = \left\{\mathbf{Z} \in \mathbb{R}^{d\times d} : \mathrm{Tr}(\mathbf{Z}) = 1, \mathbf{Z} \succeq 0\right\}$
- $\mathbf{G}$ is a symmetric matrix.
- $\mathbf{A} \bullet \mathbf{B}$ stands for the scalar product of two square matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \bullet \mathbf{B} = \sum_{i,j} a_{ij} b_{ij}$.
- The LMO is a semidefinite program itself, but of a simple form that allows an explicit solution.
- **Atoms**: The matrices of the form $\mathbf{z}\mathbf{z}^\top$ with $\mathbf{z} \in \mathbb{R}^d, \|\mathbf{z}\| = 1$ (these are positive semidefinite of trace 1 and hence in X).

Need to show: every $\mathbf{Z} \in X$ is a convex combination of atoms.

- diagonalize: $\mathbf{Z} = \mathbf{T}\mathbf{D}\mathbf{T}^\top$ where $\mathbf{T}$ is orthogonal and $\mathbf{D}$ is diagonal, of trace 1 .
- $\mathbf{D}$'s diagonal elements $\lambda_1, \dots, \lambda_d$ are the (nonnegative) eigenvalues of $\mathbf{Z}$.
- Let $\mathbf{a}_i$ be the i-th column of $\mathbf{T}$. As $\mathbf{T}$ is orthogonal, we have $\|\mathbf{a}_i\| = 1$.
- $\mathbf{Z} = \sum_{i=1}^d \lambda_i \mathbf{a}_i \mathbf{a}_i^\top$ is the desired convex combination of atoms.

Lemma 9.1

Let λ_1 be the smallest eigenvalue of $\mathbf{G}$, and let $\mathbf{s}_1$ be a corresponding eigenvector of unit length. Then we can choose $\mathrm{LMO}_X(\mathbf{G}) = \mathbf{s}_1\mathbf{s}_1^\top$.

Lemma 9.2

Duality gap: $g(\mathbf{x}) := \nabla f(\mathbf{x})^\top(\mathbf{x} - \mathbf{s})$ for $\mathbf{s} := \mathrm{LMO}_X(\nabla f(\mathbf{x}))$.
Suppose that the constrained minimization problem $\min\{f(\mathbf{x}) : \mathbf{x} \in X\}$ has a minimizer $\mathbf{x}^\star$. Let $\mathbf{x} \in X$. Then

$$g(\mathbf{x}) \ge f(\mathbf{x}) - f\left(\mathbf{x}^\star\right)$$

meaning that the duality gap is an upper bound for the optimality gap.

Thm 9.3 Convergence in $\mathcal{O}(1/\varepsilon)$ steps

Consider the constrained minimization problem $\min\{f(\mathbf{x}) : \mathbf{x} \in X\}$ where $f : \mathbb{R}^d \to \mathbb{R}$ is convex and smooth with parameter L, and set X is convex, closed and bounded (in particular, a minimizer $\mathbf{x}^\star$ of f over X exists, and all linear minimization oracles have minimizers). With any $\mathbf{x}_0 \in X$, and with stepsizes $\gamma_t = 2/(t+2)$, the Frank-Wolfe algorithm yields

$$f(\mathbf{x}_T) - f\left(\mathbf{x}^\star\right) \le \frac{2L\,\mathrm{diam}(X)^2}{T+1}, \quad T \ge 1$$

where $\mathrm{diam}(X) := \max_{\mathbf{x},\mathbf{y}\in X}\|\mathbf{x} - \mathbf{y}\|$ is the diameter of X (which exists since X is closed and bounded).

- **Standard stepzise** in the Frank-Wolfe algorithm: $\gamma_t = 2/(t+2)$.
- We need to assume that f **is smooth**, but the smoothness parameter L does not enter the stepsize.

Lemma 9.4 Descent Lemma

For a step $\mathbf{x}_{t+1} := \mathbf{x}_t + \gamma_t(\mathbf{s} - \mathbf{x}_t)$ with stepsize $\gamma_t \in [0, 1]$, it holds that

$$f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) - \gamma_t g(\mathbf{x}_t) + \gamma_t^2 \frac{L}{2}\|\mathbf{s} - \mathbf{x}_t\|^2$$

where $\mathbf{s} = \mathrm{LMO}_X(\nabla f(\mathbf{x}_t))$.

Stepsize variants

Writing $h(\mathbf{x}) := f(\mathbf{x}) - f\left(\mathbf{x}^\star\right)$ for the (unknown) optimization gap at point $\mathbf{x}$, and we have $h(\mathbf{x}) \le g(\mathbf{x})$.
Let $C := \frac{L}{2}\operatorname{diam}(X)^2$. Thm 9.3 can be written as

$$f(\mathbf{x}_t) - f\left(\mathbf{x}^\star\right) = h(\mathbf{x}_t) \le \frac{4C}{t+1}, \quad t \ge 1$$

Line search stepsize

$$\gamma_t := \underset{\gamma\in[0,1]}{\operatorname{argmin}} f\left((1-\gamma)\mathbf{x}_t + \gamma\mathbf{s}\right)$$

.
Let $\mathbf{y}_{t+1}$ be the iterate obtained from $\mathbf{x}_t$ with the standard stepsize $\mu_t = 2(t+2)$. We return to the previous analysis:

$$h(\mathbf{x}_{t+1}) \le h(\mathbf{y}_{t+1}) \le (1-\mu_t)h(\mathbf{x}_t) + \mu_t^2 C$$

Gap-based stepsize

Choose γ_t such that the term $-\gamma_t g(\mathbf{x}_t) + \gamma_t^2\frac{L}{2}\|\mathbf{s}-\mathbf{x}_t\|^2$ on the right-hand side of the inequality for $h(\mathbf{x}_{t+1})$ a is minimized.

$$\gamma_t := \min\left(\frac{g(\mathbf{x}_t)}{L\|\mathbf{s}-\mathbf{x}_t\|^2}, 1\right)$$

Affinely Equivalent

(f, X) and (f', X') are called affinely equivalent if $f'(\mathbf{x}) = f(A\mathbf{x}+\mathbf{b})$ for some invertible matrix A and some vector $\mathbf{b}$, and $X' = \left\{A^{-1}(\mathbf{x}-\mathbf{b}) : \mathbf{x} \in X\right\}$.
We have $\mathbf{x} \in X$ with function value $f(\mathbf{x})$ if and only if $\mathbf{x}' = A^{-1}(\mathbf{x}-\mathbf{b}) \in X'$ with the same function value $f'(\mathbf{x}') = f\left(AA^{-1}(\mathbf{x}-\mathbf{b})+\mathbf{b}\right) = f(\mathbf{x})$.

Affine invariance of the Frank-Wolfe algorithm

The Frank-Wolfe algorithm is invariant under all affine transformations of space.
Let (f, X) and (f', X') be affinely equivalent as before.
The points $\mathbf{x}$ and $\mathbf{x}' = \mathbf{A}^{-1}(\mathbf{x}-\mathbf{b}) \in X'$ are said to correspond to each other.
Chain rule: $\nabla f'(\mathbf{x}') = \mathbf{A}^\top \nabla f(\mathbf{A}\mathbf{x}'+\mathbf{b}) = \mathbf{A}^\top \nabla f(\mathbf{x})$.
Now consider performing an iteration of the Frank-Wolfe algorithm
(a) on (f, X), starting from some iterate $\mathbf{x}$, and
(b) on (f', X'), starting from the corresponding iterate $\mathbf{x}'$,
in both cases with the same stepsize.
Corresponding linear function values:

$$\nabla f'\left(\mathbf{x}'\right)^\top \mathbf{z}' = \nabla f(\mathbf{x})^\top \mathbf{A}\mathbf{A}^{-1}(\mathbf{z}-\mathbf{b}) = \nabla f(\mathbf{x})^\top \mathbf{z} - c$$

where c some constant.
Corresponding steps: $\mathbf{s} = \mathrm{LMO}_X(\nabla f(\mathbf{x}))$ if and only if $\mathbf{s}' = \mathrm{LMO}_{X'}(\nabla f'(\mathbf{x}'))$

Curvature Constant

Curvature constant (notion of complexity of (f, X)):

$$C_{(f,X)} := \sup_{\substack{\mathbf{x},\mathbf{s}\in X, \gamma\in(0,1] \\ \mathbf{y}=(1-\gamma)\mathbf{x}+\gamma\mathbf{s}}} \frac{1}{\gamma^2}\left(f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x})\right).$$

The curvature constant is affine invariant, i.e. if (f, X) and (f', X') are affinely equivalent, then $C_{(f,X)} = C_{(f',X')}$.

Thm 9.5 Convergence in terms of the curvature constant

Consider the constrained minimization problem where $f : \mathbb{R}^d \to \mathbb{R}$ is convex, and set X is convex, closed and bounded. Let $C_{(f,X)}$ be the curvature constant of f over X. With $\mathbf{x}_0 \in X$, and with stepsizes $\gamma_t = 2/(t+2)$, the Frank-Wolfe algorithm yields

$$f(\mathbf{x}_T) - f\left(\mathbf{x}^\star\right) \le \frac{4C_{(f,X)}}{T+1}, \quad T \ge 1$$

$\mathcal{O}(1/\varepsilon)$ many iterations are sufficent to obtain optimality gap at most ε.

Lemma 9.6 Relating Curvature and Smoothness

Let f be a convex function which is smooth with parameter L over X. Then

$$C_{(f,X)} \le \frac{L}{2}\operatorname{diam}(X)^2$$

Convergence in duality gap

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and smooth with parameter L, and $\mathbf{x}_0 \in X, T \ge 2$. Then choosing any of the three stepsizes that we have discussed, the Frank-Wolfe algorithm guarantees some $t, 1 \le t \le T$ such that

$$g(\mathbf{x}_t) \le \frac{27/2 \cdot C_{(f,X)}}{T+1}, \quad T \ge 2$$

The smallest value $g(\mathbf{x}_t), t = 1, \ldots, T$ bounds the optimality gap at iteration t :

$$f(\mathbf{x}_t) - f\left(\mathbf{x}^\star\right) \le g(\mathbf{x}_t) \le \frac{27/2 \cdot C_{(f,X)}}{T+1}$$

This is a computable bound that certifies small optimality gap.

Coreset

The current solution is a convex combination of $\mathbf{x}_0$ and $\mathcal{O}(1/\varepsilon)$ many extreme points (atoms) of the constraint set X.
Thinking of ε as a constant (such as 0.01): constantly many extreme points are sufficient in order to get an almost optimal solution.
Coreset: a small subsets of a given set of objects that is representative (with respect to some measure) for the set of all objects.
Some algorithms for finding small coresets are variants of or inspired by the Frank-Wolfe algorithm

10. Newton's Method and Quasi-Newton Methods

Newton-Raphson Method: 1-dim Case

Goal: find a zero of differentiable $f : \mathbb{R} \to \mathbb{R}$.
Method:

$$x_{t+1} := x_t - \frac{f(x_t)}{f'(x_t)}, \quad t \ge 0.$$

The Babylonian Method (Computing square roots)

Computing square roots: find a zero of $f(x) = x^2 - R, R \in \mathbb{R}_+$.
Newton-Raphson step:

$$x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)} = x_t - \frac{x_t^2 - R}{2x_t} = \frac{1}{2}\left(x_t + \frac{R}{x_t}\right).$$

Starting from $x_0 > 0$, we have

$$x_{t+1} = \frac{1}{2}\left(x_t + \frac{R}{x_t}\right) \ge \frac{x_t}{2}.$$

Starting from $x_0 = R \ge 1$, it takes at least $\log(R)/2$ steps to get $x_t < 2\sqrt{R}$.
But still only $\mathcal{O}(\log R)$ steps to get $x_t - \sqrt{R} < 1/2$. (**Proof see Hw10 Ex1**)
Suppose $x_0 - \sqrt{R} < 1/2$ (achievable after $\mathcal{O}(\log R)$ steps).

$$x_{t+1} - \sqrt{R} = \frac{1}{2}\left(x_t + \frac{R}{x_t}\right) - \sqrt{R} = \frac{1}{2x_t}\left(x_t - \sqrt{R}\right)^2$$

Assume $R \ge 1/4$. Then all iterates have value at least $\sqrt{R} \ge 1/2$. Hence we get

$$x_{t+1} - \sqrt{R} \le \left(x_t - \sqrt{R}\right)^2$$

$$x_T - \sqrt{R} \le \left(x_0 - \sqrt{R}\right)^{2^T} < \left(\frac{1}{2}\right)^{2^T}, \quad T \ge 0.$$

To get $x_T - \sqrt{R} < \varepsilon$, we only need $T = \log\log\left(\frac{1}{\varepsilon}\right)$ steps!

Newton's method for optimization

1-dimensional case

Find a global minimum $x^\star$ of a twice-differentiable convex function $f : \mathbb{R} \to \mathbb{R}$.
Update step:

$$x_{t+1} := x_t - \frac{f'(x_t)}{f''(x_t)} = x_t - f''(x_t)^{-1} f'(x_t)$$

d-dimensional case

Newton's method for minimizing a convex function $f : \mathbb{R}^d \to \mathbb{R}$:

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t)$$

Adaptive gradient descent

General update scheme:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \mathbf{H}(\mathbf{x}_t)\nabla f(\mathbf{x}_t),$$

where $\mathbf{H}(\mathbf{x}) \in \mathbb{R}^{d\times d}$ is some matrix.
Newton's method: $\mathbf{H} = \nabla^2 f(\mathbf{x}_t)^{-1}$.
Gradient descent: $\mathbf{H} = \gamma\mathbf{I}$.
Newton's method: adaptive gradient descent, adaptation is w.r.t. the local geometry of the function at $\mathbf{x}_t$, as captured by the Hessian $\nabla^2 f(\mathbf{x}_t)$.

Nondegenerate quadratic function

A nondegenerate quadratic function is a function of the form

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{M}\mathbf{x} - \mathbf{q}^\top\mathbf{x} + c$$

where $\mathbf{M} \in \mathbb{R}^{d\times d}$ is an invertible symmetric matrix, $\mathbf{q} \in \mathbb{R}^d, c \in R$.
Let $\mathbf{x}^\star = \mathbf{M}^{-1}\mathbf{q}$ be the unique solution of $\nabla f(\mathbf{x}) = \mathbf{0}$.
$\mathbf{x}^\star$ is the unique global minimum if f is convex.

Lemma 10.1 Convergence in one step on quadratic functions

On nondegenerate quadratic functions, with any starting point $\mathbf{x}_0 \in \mathbb{R}^d$, Newton's method yields $\mathbf{x}_1 = \mathbf{x}^\star$.

Lemma 10.3 Minimizing the second-order Taylor approximation

Let f be convex and twice differentiable at $\mathbf{x}_t \in \operatorname{dom}(f)$, with $\nabla^2 f(\mathbf{x}_t) \succ 0$ being invertible. The vector $\mathbf{x}_{t+1}$ resulting from the Netwon step satisfies

$$\mathbf{x}_{t+1} = \underset{\mathbf{x}\in\mathbb{R}^d}{\operatorname{argmin}}\; f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top(\mathbf{x}-\mathbf{x}_t) + \frac{1}{2}(\mathbf{x}-\mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{x}-\mathbf{x}_t)$$

Alternative interpretation of Newton's method: Each step minimizes the local second-order Taylor approximation.

Thm 10.4 Convergence Thm

Let $f : \operatorname{dom}(f) \to \mathbb{R}$ be twice differentiable with a critical point $\mathbf{x}^\star$. Suppose there is a ball $X \subseteq \operatorname{dom}(f)$ with center $\mathbf{x}^\star$, s.t.
(i) Bounded inverse Hessians: There exists a real number $\mu > 0$ such that

$$\left\|\nabla^2 f(\mathbf{x})^{-1}\right\| \le \frac{1}{\mu}, \quad \forall \mathbf{x} \in X.$$

(ii) Lipschitz continuous Hessians: There exists a real number $B > 0$ such that

$$\left\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\right\| \le B\|\mathbf{x}-\mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in X.$$

Then, for $\mathbf{x}_t \in X$ and $\mathbf{x}_{t+1}$ resulting from the Newton step, we have

$$\left\|\mathbf{x}_{t+1} - \mathbf{x}^\star\right\| \le \frac{B}{2\mu}\left\|\mathbf{x}_t - \mathbf{x}^\star\right\|^2$$

Corollary 10.5 Super-exponentially fast

With the assumptions and terminology of the convergence theorem, and if

$$\left\|\mathbf{x}_0 - \mathbf{x}^\star\right\| \le \frac{\mu}{B}$$

then Newton's method yields

$$\left\|\mathbf{x}_T - \mathbf{x}^\star\right\| \le \frac{\mu}{B}\left(\frac{1}{2}\right)^{2^T-1}, \quad T \ge 0$$

Starting close to the critical point, we will reach distance at most ε to it within $\mathcal{O}(\log\log(1/\varepsilon))$ steps.

Lemma 10.6

With the assumptions and terminology of the convergence theorem, and if $\mathbf{x}_0 \in X$ satisfies

$$\left\|\mathbf{x}_0 - \mathbf{x}^\star\right\| \le \frac{\mu}{B},$$

Then the Hessians in Newton's method satisfy the relative error bound

$$\frac{\left\|\nabla^2 f(\mathbf{x}_t) - \nabla^2 f\left(\mathbf{x}^\star\right)\right\|}{\left\|\nabla^2 f(\mathbf{x}^\star)\right\|} \le \left(\frac{1}{2}\right)^{2^t-1}, \quad t \ge 0.$$

Lemma 10.7 Strong convexity ⇒ Bounded inverse Hessians

Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be twice differentiable and strongly convex with parameter μ over an open convex subset $X \subseteq \mathbf{dom}(f)$ meaning that

$$f(\mathbf{y}) \ge f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2, \quad \forall \mathbf{x},\mathbf{y} \in X$$

Then $\nabla^2 f(\mathbf{x})$ is invertible and $\left\|\nabla^2 f(\mathbf{x})^{-1}\right\| \le 1/\mu$ for all $\mathbf{x} \in X$, where $\|\cdot\|$ is the spectral norm.

Secant Method: 1-dim Case

Use finite difference approximation of $f'(x_t)$:

$$f'(x_t) \approx \frac{f(x_t) - f(x_{t-1})}{x_t - x_{t-1}}.$$

(for $|x_t - x_{t-1}|$ small)
Obtain the **secant method**:

$$x_{t+1} := x_t - f(x_t)\frac{x_t - x_{t-1}}{f(x_t) - f(x_{t-1})}$$

Secant condition

Apply finite difference approximation to f'' (still **1-dim**), $\Leftrightarrow$

$$H_t := \frac{f'(x_t) - f'(x_{t-1})}{x_t - x_{t-1}} \approx f''(x_t)$$
$$f'(x_t) - f'(x_{t-1}) = H_t(x_t - x_{t-1})$$

the **secant condition**.
Newton's method: $x_{t+1} := x_t - f''(x_t)^{-1} f'(x_t)$
Secant method: $x_{t+1} := x_t - H_t^{-1} f'(x_t)$
In higher dimensions: Let $H_t \in \mathbb{R}^{d\times d}$ be an invertible symmetric matrix satisfying the **d-dimensional secant condition**

$$\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}) = H_t(\mathbf{x}_t - \mathbf{x}_{t-1}).$$

The secant method step then becomes

$$\mathbf{x}_{t+1} := \mathbf{x}_t - H_t^{-1}\nabla f(\mathbf{x}_t).$$

Quasi-Newton Methods

The secant method approximates Newton's method.

- $d = 1$: unique number H_t satisfying the secant condition
- $d > 1$: Secant condition $\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}) = H_t(\mathbf{x}_t - \mathbf{x}_{t-1})$ has infinitely many symmetric solutions H_t (underdetermined linear system).

Any scheme of choosing in each step of the secant method a **symmetric** H_t that satisfies the secant condition defines a **Quasi Newton method**.

Greenstadt's family of Quasi-Newton methods

Given: iterates $\mathbf{x}_{t-1}, \mathbf{x}_t$ as well as the matrix $\mathbf{H}_{t-1}^{-1}$.
Wanted: next matrix $\mathbf{H}_t^{-1}$ needed in next Quasi-Newton step

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \mathbf{H}_t^{-1}\nabla f(\mathbf{x}_t).$$

Greenstadt: Update

$$\mathbf{H}_t^{-1} := \mathbf{H}_{t-1}^{-1} + \mathbf{E}_t,$$

$\mathbf{E}_t$ an error matrix. Try to minimize the errror subject to $\mathbf{H}_t$ satisfying the secant condition! Simple error measure: squared Frobenius norm

$$\|\mathbf{E}\|_F^2 := \sum_{i=1}^{d}\sum_{j=1}^{d} E_{ij}^2.$$

More general error measure

$$\left\|\mathbf{A}\,\mathbf{E}\,\mathbf{A}^\top\right\|_F^2,$$

where $\mathbf{A} \in \mathbb{R}^{d\times d}$ is some fixed invertible transformation matrix. $\mathbf{A} = \mathbf{I}$: squared Frobenius norm of $\mathbf{E}$, the ßpecialized"method.

The Greenstadt Update $H_{t-1}^{-1} \to H_t^{-1}$

Secant condition in terms of H_t^{-1} :

$$H_t^{-1}(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})) = (\mathbf{x}_t - \mathbf{x}_{t-1}).$$

Fix t and simplify notation:

$$\begin{aligned}
\mathbf{H} &:= \mathbf{H}_{t-1}^{-1} && \text{(old inverse)}\\
\mathbf{H}' &:= \mathbf{H}_t^{-1} && \text{(new inverse)}\\
\mathbf{E} &:= \mathbf{E}_t, && \text{(error matrix)}\\
\boldsymbol{\sigma} &:= \mathbf{x}_t - \mathbf{x}_{t-1} && \text{(step in solutions)}\\
\mathbf{y} &= \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}) && \text{(step in gradients)}\\
\mathbf{r} &= \boldsymbol{\sigma} - H\mathbf{y}
\end{aligned}$$

The update formula is

$$\mathbf{H}' = \mathbf{H} + \mathbf{E},$$

Secant condition becomes

$$\begin{aligned}
&\mathbf{H}'\mathbf{y} = \boldsymbol{\sigma} \quad\Leftrightarrow\quad (\mathbf{H}+\mathbf{E})\mathbf{y} = \boldsymbol{\sigma}\\
\Leftrightarrow\quad &\mathbf{E}\mathbf{y} = \boldsymbol{\sigma} - \mathbf{H}\mathbf{y} \quad\Leftrightarrow\quad \mathbf{E}\mathbf{y} = \mathbf{r}
\end{aligned}$$

Minimizing the error becomes a convex constrained minimization problem in the d^2 variables E_{ij}:

$$\begin{array}{lll}
\text{minimize} & \frac{1}{2}\left\|\mathbf{A}\,\mathbf{E}\,\mathbf{A}^\top\right\|_F^2 & \text{(error function)}\\
\text{subject to} & \mathbf{E}\,\mathbf{y} = \mathbf{r} & \text{(secant condition)}\\
 & \mathbf{E}^\top = \mathbf{E} & \text{(symmetry)}
\end{array}$$

Don't need to solve it computationally (for numbers E_{ij}), but mathematically (formula for $\mathbb{E}$).
Minimize convex quadratic function subject to linear equations → analytic formula for the minimizer from the **method of Lagrange multipliers**.

Thm 10.8 (11.1) The method of Lagrange multipliers

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable, $\mathbf{C} \in \mathbb{R}^{m\times d}$ for some $m \in \mathbb{N}$, $\mathbf{e} \in \mathbb{R}^m$, $\mathbf{x}^\star \in \mathbb{R}^d$ such that $\mathbf{C}\mathbf{x}^\star = \mathbf{e}$. Then the following two statements are equivalent.
(i) $\mathbf{x}^\star = \operatorname{argmin}\left\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^d, \mathbf{C}\mathbf{x} = \mathbf{e}\right\}$
(ii) There exists a vector $\lambda \in \mathbb{R}^m$ such that

$$\nabla f\left(\mathbf{x}^\star\right)^\top = \lambda^\top\mathbf{C}$$

The entries of λ are known as the **Lagrange multipliers**.

Greenstadt method

Let $\mathbf{M} \in \mathbb{R}^{d\times d}$ be a symmetric and invertible matrix. Consider the Quasi-Newton method

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \mathbf{H}_t^{-1}\nabla f(\mathbf{x}_t), \quad t \ge 1,$$

where $\mathbf{H}_0 = \mathbf{I}$ (or any positive definite matrix), and $\mathbf{H}_t^{-1} = \mathbf{H}_{t-1}^{-1} + \mathbf{E}_t$ for all $t \ge 1$. For any fixed t, set $\mathbf{H} := \mathbf{H}_{t-1}^{-1}, \mathbf{H}' := \mathbf{H}_t^{-1}, \boldsymbol{\sigma} := \mathbf{x}_t - \mathbf{x}_{t-1}, \mathbf{y} := \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})$.
Define

$$\begin{aligned}
\mathbf{E}^\star = &\frac{1}{\mathbf{y}^\top\mathbf{M}\mathbf{y}}\left(\boldsymbol{\sigma}\mathbf{y}^\top\mathbf{M} + \mathbf{M}\mathbf{y}\boldsymbol{\sigma}^\top - \mathbf{H}\mathbf{y}\mathbf{y}^\top\mathbf{M} - \mathbf{M}\mathbf{y}\mathbf{y}^\top\mathbf{H}\right.\\
&\left.-\frac{1}{\mathbf{y}^\top\mathbf{M}\mathbf{y}}\left(\mathbf{y}^\top\boldsymbol{\sigma} - \mathbf{y}^\top\mathbf{H}\mathbf{y}\right)\mathbf{M}\mathbf{y}\mathbf{y}^\top\mathbf{M}\right)
\end{aligned}$$

If the update matrix $\mathbf{E}_t = \mathbb{E}^\star$ is used, the method is called the **Greenstadt method** with parameter $\mathbf{M}$.
Greenstadt suggested

$$\begin{aligned}
\mathbf{M} &= \mathbf{I} \quad \text{(default choice)}\\
\mathbf{M} &= \mathbf{H} \quad \left(\text{previous inverse } \mathbf{H}_{t-1}^{-1}\right)
\end{aligned}$$

BFGS method

Chose $\mathbf{M} = \mathbf{H}'$. Secant condition holds: $\mathbf{M}\mathbf{y} = \mathbf{H}'\mathbf{y} = \boldsymbol{\sigma}$. $\mathbf{M}$ cancels.
The **BFGS method** is the Greenstadt method with parameter $\mathbf{M} := \mathbf{H}' = \mathbf{H}_t^{-1}$ in step t, in which case the update matrix $\mathbf{E}^\star$ assumes the form

$$\begin{aligned}
\mathbf{E}^\star = &\frac{1}{\mathbf{y}^\top\boldsymbol{\sigma}}\left(2\boldsymbol{\sigma}\boldsymbol{\sigma}^\top - \mathbf{H}\mathbf{y}\boldsymbol{\sigma}^\top - \boldsymbol{\sigma}\mathbf{y}^\top\mathbf{H}\right.\\
&\left.-\frac{1}{\boldsymbol{\sigma}^\top\mathbf{y}}\left(\mathbf{y}^\top\boldsymbol{\sigma} - \mathbf{y}^\top\mathbf{H}\mathbf{y}\right)\boldsymbol{\sigma}^\top\right)\\
= &\frac{1}{\mathbf{y}^\top\boldsymbol{\sigma}}\left(-\mathbf{H}\mathbf{y}\boldsymbol{\sigma}^\top - \boldsymbol{\sigma}\mathbf{y}^\top\mathbf{H} + \left(1 + \frac{\mathbf{y}^\top\mathbf{H}\mathbf{y}}{\mathbf{y}^\top\boldsymbol{\sigma}}\right)\boldsymbol{\sigma}\boldsymbol{\sigma}^\top\right)
\end{aligned}$$

where $\mathbf{H} = \mathbf{H}_{t-1}^{-1}, \boldsymbol{\sigma} = \mathbf{x}_t - \mathbf{x}_{t-1}, \mathbf{y} = \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})$.

$$\begin{aligned}
\mathbf{H}' &= \mathbf{H} + \mathbf{E}^*\\
&= \mathbf{H} + \frac{1}{\mathbf{y}^\top\boldsymbol{\sigma}}\left\{-\mathbf{H}\mathbf{y}\boldsymbol{\sigma}^\top - \boldsymbol{\sigma}\mathbf{y}^\top\mathbf{H} + \left(1 + \frac{\mathbf{y}^\top\mathbf{H}\mathbf{y}}{\mathbf{y}^\top\boldsymbol{\sigma}}\right)\boldsymbol{\sigma}\boldsymbol{\sigma}^\top\right\}\\
&= \left(\mathbf{I} - \frac{\boldsymbol{\sigma}\mathbf{y}^\top}{\mathbf{y}^\top\boldsymbol{\sigma}}\right)\mathbf{H}\left(\mathbf{I} - \frac{\mathbf{y}\boldsymbol{\sigma}^\top}{\mathbf{y}^\top\boldsymbol{\sigma}}\right) + \frac{\boldsymbol{\sigma}\boldsymbol{\sigma}^\top}{\mathbf{y}^\top\boldsymbol{\sigma}}
\end{aligned}$$

- $\mathbf{y}^\top\boldsymbol{\sigma} > 0$ unless $\mathbf{x}_{t-1} = \mathbf{x}_t$ or $f(\lambda\mathbf{x}_t + (1-\lambda)\mathbf{x}_{t-1}) = \lambda f(\mathbf{x}_t) + (1-\lambda)f(\mathbf{x}_{t-1})$ for all $\lambda \in (0,1)$. (**Proof see Hw10 Ex4(i)**)
- If $\mathbf{y}^\top\boldsymbol{\sigma} > 0$ and $\mathbf{H}$ is positive definite, then also $\mathbf{H}'$ is positive definite. In this respect, the matrices in the BFGS method behave like proper inverse Hessians. (**Proof see Hw10 Ex4(ii)**)
- Cost per update step: $\mathcal{O}\left(d^2\right)$, no Hessians and no inversions required.
- Newton and Quasi-Newton methods are often performed with **scaled steps**. This means that the iteration becomes

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha_t H_t^{-1}\nabla f(\mathbf{x}_t), \quad t \ge 1,$$

for some $\alpha_t \in \mathbb{R}_+$.

Observation 10.9 (11.6)

With $\mathbf{E}^\star$ as in the BFGS method and $\mathbf{H}' = \mathbf{H} + \mathbf{E}^\star$, we have

$$\mathbf{H}'$$

Lemma 10.10 (11.7) Efficiently computing matrix vector-products $\mathbf{H}_t^{-1}\nabla f(\mathbf{x}_t)$

Let $\mathbf{H} = \mathbf{H}_{t-1}^{-1}, \mathbf{H}' = \mathbf{H}_t^{-1}$ as in the BFGS method, i.e.

$$\mathbf{H}' = \left(\mathbf{I} - \frac{\sigma \mathbf{y}^\top}{\mathbf{y}^\top \sigma}\right)\mathbf{H}\left(\mathbf{I} - \frac{\mathbf{y}\sigma^\top}{\mathbf{y}^\top \sigma}\right) + \frac{\sigma\sigma^\top}{\mathbf{y}^\top \sigma}.$$

Let $\mathbf{g}' \in \mathbb{R}^d$. Suppose that we have an oracle to compute $\mathbf{s} = \mathbf{H}\mathbf{g}$ for any vector $\mathbf{g}$. Then $\mathbf{s}' = \mathbf{H}'\mathbf{g}'$ can be computed with **one oracle call and $\mathcal{O}(d)$ additional arithmetic operations**, assuming that σ and $\mathbf{y}$ are known.

The recursive BFGS-step (update step in one iteration)

Handout10 Page 44
In iteration t, call BFGS-STEP$(t, \nabla f(\mathbf{x}_t))$ to get $H_t^{-1}\nabla f(\mathbf{x}_t)$.
Runtime $\mathcal{O}(td)$.
Worse than before if $t > d$.

The L-BFGS method (update step in one iteration)

Handout10 Page 46
In iteration t, call L-BFGS-STEP$(t, m, \nabla f(\mathbf{x}_t))$ to get an approximation of $\mathbf{H}_t^{-1}\nabla f(\mathbf{x}_t)$ based on the previous m iterations.
Runtime per update $\mathcal{O}(dm) = \mathcal{O}(d)$ if m is constant.

11. Modern Second-Order Methods and Nonconvex Optimization

Part A: Modern Second-order Methods

Global analysis for strongly-convex smooth objectives

- Assume $f(\mathbf{x})$ is μ-strongly convex and has L-Lipschitz continuous gradient.
- Consider Newton method $\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma\nabla^2 f(\mathbf{x}_t)^{-1}\nabla f(\mathbf{x}_t)$.
- We can show that Newton method enjoys a globally linear convergence with properly chosen stepsize $\gamma > 0$:

$$f(\mathbf{x}_t) - f^* \le \left(1 - \frac{\mu^2}{L^2}\right)^t (f(\mathbf{x}_0) - f^*)$$

- Note that this is **worse than GD**, where

$$f(\mathbf{x}_t) - f^* \le \left(1 - \frac{\mu}{L}\right)^t (f(\mathbf{x}_0) - f^*)$$

Overcoming the local nature of Newton method

- **Newton method with line-search**: select γ_t such that $f(\mathbf{x}_{t+1}) < f(\mathbf{x}_t)$ with sufficient decrease.

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t \nabla^2 f(\mathbf{x}_t)^{-1}\nabla f(\mathbf{x}_t)$$

- **Damped Newton method**:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{1 + \lambda_f(\mathbf{x}_t)}\nabla^2 f(\mathbf{x}_t)^{-1}\nabla f(\mathbf{x}_t),$$

  where $\lambda_f(\mathbf{x}) = \left\|\left[\nabla^2 f(\mathbf{x})\right]^{-1/2}\nabla f(\mathbf{x})\right\|$ is the Newton decrement.

- **Regularization approach**: regularize the Hessian and adjust γ_t properly

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \left[\gamma_t I + \nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t)$$

- **Trust-region approach**:

$$\begin{aligned}\mathbf{x}_{t+1} = \underset{\mathbf{x}}{\operatorname{argmin}} \quad & f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top(\mathbf{x} - \mathbf{x}_t) \\ & + \frac{1}{2}(\mathbf{x} - \mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{x} - \mathbf{x}_t) \\ \text{s.t. } & \|\mathbf{x} - \mathbf{x}_t\| \le \Delta_k.\end{aligned}$$

Lipschitz Hessian

- Recall for functions f with L_1-Lipschitz gradient, we have

$$f(\mathbf{x}) \le f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top(\mathbf{x} - \mathbf{x}_t) + \frac{L_1}{2}\|\mathbf{x} - \mathbf{x}_t\|^2$$

  GD can be viewed as iteratively minimizing the quadratic upper bound function.

- Now assuming f **has L_2-Lipschitz Hessian**, similarly, we can show that

$$\begin{aligned}f(\mathbf{x}) \le\ & f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top(\mathbf{x} - \mathbf{x}_t) \\ & + \frac{1}{2}(\mathbf{x} - \mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{x} - \mathbf{x}_t) + \frac{L_2}{6}\|\mathbf{x} - \mathbf{x}_t\|^3\end{aligned}$$

(**Proof see Hw11 Ex4**)

Cubic Regularization: The Algorithm

$$\mathbf{x}_{t+1} \in \underset{\mathbf{x}}{\operatorname{argmin}} \hat{f}(\mathbf{x}, \mathbf{x}_t)$$

$$\begin{aligned}\hat{f}(\mathbf{x}, \mathbf{x}_t) :=\ & f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top(\mathbf{x} - \mathbf{x}_t) \\ & + \frac{1}{2}(\mathbf{x} - \mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{x} - \mathbf{x}_t) + \frac{M}{6}\|\mathbf{x} - \mathbf{x}_t\|^3\end{aligned}$$

Cubic Regularization: Global Analysis

Key facts:

- $\nabla^2 f(\mathbf{x}_t) + \frac{M}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\| \cdot I \succeq 0$
- $\|\nabla f(\mathbf{x}_{t+1})\| \le \frac{L_2 + M}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2$
- $-f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) \ge \frac{M}{12}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^3$ if $M \ge L_2$.

Implications:

- **Convergence to a second-order stationary point**: If $\mathbf{x}^*$ is a limiting point, then $\nabla f(\mathbf{x}^*) = 0, \nabla^2 f(\mathbf{x}^*) \succeq 0$
- **Convergence rate**: We have $\min_{1 \le i \le t}\|\nabla f(\mathbf{x}_i)\| = \mathcal{O}\left(\frac{1}{t^{2/3}}\right)$.
- **Convex setting**: If f is convex, we have $f(\mathbf{x}_t) - f^* = \mathcal{O}\left(\frac{1}{t^2}\right)$.
- **Strongly convex setting**: If f is strongly convex, it implies superlinear convergence.

(**Proof see Handout11 Page 14**)

Cubic Regularization: Extension

- **Accelerated Cubic Regularization**: For convex functions, can achieve $\mathcal{O}\left(\frac{1}{t^3}\right)$ convergence rate
- **High-order Tensor Method**: For convex functions and p-th order method, can achieve $\mathcal{O}\left(\frac{1}{t^{p+1}}\right)$ convergence rate.
- **Lower bound**: For p-th order method, the lower complexity bound is $\Omega\left(\frac{1}{t^{(3p+1)/2}}\right)$

Part B: Modern Nonconvex Optimization

Nonconvex SGD

Consider the stochastic optimization problem:

$$\min_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x}) := \mathbb{E}_\xi[f(\mathbf{x}, \xi)] \quad \left[\text{or } F(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})\right]$$

SGD:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t \nabla f(\mathbf{x}_t, \xi_t), \text{ where } \xi_t \overset{iid}{\sim} P(\xi)$$

Thm 11.1 Convergence of Nonconvex SGD

Assume that

- Function F is L-smooth,
- The unbiased estimator satisfies that for all $\mathbf{x}$: $\mathbb{E}\left[\|\nabla f(\mathbf{x}, \xi) - \nabla F(\mathbf{x})\|_2^2\right] \le \sigma^2$.

Under the above assumption, SGD with $\gamma_t = \min\left\{\frac{1}{L}, \frac{\gamma}{\sigma\sqrt{T}}\right\}$ achieves:

$$\mathbb{E}\left[\|\nabla F(\hat{\mathbf{x}}_T)\|^2\right] \le \frac{\sigma}{\sqrt{T}}\left(\frac{2(F(\mathbf{x}_1) - F(\mathbf{x}^*))}{\gamma} + L\gamma\right)$$

where $\hat{\mathbf{x}}_T$ is selected uniformly at random from $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$.

Variance-reduced SGD for Nonconvex Optimization

Handout11 Pages 29-30

Stationary Points

A **stationary point** can be a local minimum, a local maximum, or a saddle point.

- If $\nabla^2 F(\bar{\mathbf{x}}) \succ 0$, then $\bar{\mathbf{x}}$ is a **local minimum**.
- If $\nabla^2 F(\bar{\mathbf{x}}) \prec 0$, then $\bar{\mathbf{x}}$ is a **local maximum**.
- If $\nabla^2 F(\bar{\mathbf{x}})$ has positive and negative eigenvalues, then $\bar{\mathbf{x}}$ is a **(strict) saddle point**.
- Otherwise, it remains inconclusive.

Thm 11.2 GD with Random Initialization (informal)

If f satisfies the strict saddle property, then **GD with random initialization** and sufficiently small constant stepsize converges to a local minimum or negative infinity almost surely.
The analysis is not specific to GD and can apply to other algorithms.
*Noisy SGD

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\nabla f(\mathbf{x}_t, \xi_t) + \mathbf{z}$$

where the noise $\mathbf{z}$ is uniformly sampled from unit sphere.

Thm 11.3 (informal)

If f satisfies the strict saddle property and has Lipschitz Hessian, then Noisy SGD with sufficiently small stepsize converges to an ϵ-second order stationary point in $\mathrm{poly}(d/\epsilon)$ steps.
Handout11 Page 41

Convergence to global minima

For problems with benign nonconvexity.
Handout11 Page 44

12. Modern Nonsmooth Optimization

Part A: Mirror Descent

Bregman Divergence

Let $\omega(\cdot): \Omega \to \mathbb{R}$ be continuously differentiable on Ω and **1-strongly convex** w.r.t. some norm $\|\cdot\|$: $\omega(\mathbf{x}) \ge \omega(\mathbf{y}) + \nabla\omega(\mathbf{y})^\top(\mathbf{x} - \mathbf{y}) + \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2, \forall x, y \in \Omega$.
The **Bregman divergence** is defined as

$$V_\omega(\mathbf{x}, \mathbf{y}) = \omega(\mathbf{x}) - \omega(\mathbf{y}) - \nabla\omega(\mathbf{y})^\top(\mathbf{x} - \mathbf{y}), \forall \mathbf{x}, \mathbf{y} \in \Omega$$

which implies

$$V_\omega(\mathbf{x}, \mathbf{y}) \ge \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2$$

Examples

- **Euclidean distance**: $\Omega = \mathbb{R}^d, \omega(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2, \|\cdot\| = \|\cdot\|_2$

$$V_\omega(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$$

- **Mahalanobis distance**: $\Omega = \mathbb{R}^d, \omega(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x}$ (where $\mathbf{Q} \succeq \mathbf{I}$), $\|\cdot\| = \|\cdot\|_2$,

$$V_\omega(\mathbf{x}, \mathbf{y}) = \frac{1}{2}(\mathbf{x} - \mathbf{y})^\top\mathbf{Q}(\mathbf{x} - \mathbf{y})$$

- **Kullback-Leibler divergence**: $\Omega = \Delta_d, \omega(\mathbf{x}) = \sum_{i=1}^d x_i \log x_i, \|\cdot\| = \|\cdot\|_1$,

$$V_\omega(\mathbf{x},\mathbf{y}) = \mathrm{KL}(\mathbf{x} \mid \mathbf{y}) := \sum_{i=1}^d x_i \log \frac{x_i}{y_i}$$

Properties

- Nonnegativity: $V_\omega(\mathbf{x},\mathbf{y}) \geq 0$.
- Convexity: $V_\omega(\mathbf{x},\mathbf{y})$ is convex in $\mathbf{x}$.
- $V_\omega(\mathbf{x},\mathbf{y}) \geq \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2$.
- Non-symmetry: $V_\omega(\mathbf{x},\mathbf{y}) \neq V_\omega(\mathbf{y},\mathbf{x})$ in general.
- Generalized Pythagorean Theorem: Define the Bregman projection of a point $\mathbf{z}$ onto X as:

$$\Pi_X^\omega(\mathbf{z}) := \operatorname*{argmin}_{\mathbf{x}\in X} V_\omega(\mathbf{x},\mathbf{z}).$$

Then for any $\mathbf{x} \in X, \mathbf{z} \in \Omega$ it holds that

$$V_\omega(\mathbf{x},\mathbf{z}) \geq V_\omega\left(\mathbf{x}, \Pi_X^\omega(\mathbf{z})\right) + V_\omega\left(\Pi_X^\omega(\mathbf{z}), \mathbf{z}\right)$$

(Proof see Hw12 Ex1)

Lemma 12.1 Three Point Identity

$$V_\omega(\mathbf{x},\mathbf{z}) = V_\omega(\mathbf{x},\mathbf{y}) + V_\omega(\mathbf{y},\mathbf{z}) - \langle \nabla\omega(\mathbf{z}) - \nabla\omega(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \quad \forall \mathbf{x},\mathbf{y},\mathbf{z} \in \Omega$$

Special case: $\omega(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$, this is the law of cosine:

$$\|\mathbf{x}-\mathbf{z}\|_2^2 = \|\mathbf{x}-\mathbf{y}\|_2^2 + \|\mathbf{y}-\mathbf{z}\|_2^2 - 2\langle \mathbf{z}-\mathbf{y}, \mathbf{x}-\mathbf{y}\rangle$$

Corollary of Lemma 12.1 Three Point Identity

Since $\mathbf{x}_{t+1} = \operatorname{argmin}_{\mathbf{x}\in X}\{V_\omega(\mathbf{x},\mathbf{x}_t) + \langle \gamma_t \mathbf{g}_t, \mathbf{x}\rangle\}$, by the optimality condition,

$$\langle \nabla\omega(\mathbf{x}_{t+1}) + \gamma_t \mathbf{g}_t - \nabla\omega(\mathbf{x}_t), \mathbf{x}-\mathbf{x}_{t+1}\rangle \geq 0, \forall \mathbf{x} \in X$$
$$\Rightarrow \langle \nabla\omega(\mathbf{x}_{t+1}) - \nabla w(\mathbf{x}_t), \mathbf{x}-\mathbf{x}_{t+1}\rangle \geq \langle \gamma_{t}, \mathbf{x}_{t+1}-\mathbf{x}\rangle$$

From three point identity, we have for $\forall \mathbf{x} \in X$:

$$\begin{aligned}\langle \gamma_t \mathbf{g}_t, \mathbf{x}_{t+1}-\mathbf{x}\rangle &\leq \langle \nabla(\mathbf{x}_{t+1}) - \nabla\omega(\mathbf{x}_t), \mathbf{x}-\mathbf{x}_{t+1}\rangle \\ &= V_\omega(\mathbf{x},\mathbf{x}_t) - V_\omega(\mathbf{x},\mathbf{x}_{t+1}) - \underbrace{V_\omega(\mathbf{x}_{t+1},\mathbf{x}_t)}_{\geq \frac{1}{2}\|\mathbf{x}_t-\mathbf{x}_{t+1}\|^2}\end{aligned}$$

Mirror Descent Algorithm

$$\mathbf{x}_{t+1} = \operatorname*{argmin}_{\mathbf{x}\in X}\{V_\omega(\mathbf{x},\mathbf{x}_t) + \langle \gamma_t \mathbf{g}_t, \mathbf{x}\rangle\}, \text{ where } \mathbf{g}_t \in \partial f(\mathbf{x}_t)$$

Examples

- Subgradient descent: $\omega(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2, V_\omega(\mathbf{x},\mathbf{x}_t) = \frac{1}{2}\|\mathbf{x}-\mathbf{x}_t\|_2^2$

$$\mathbf{x}_{t+1} = \Pi_X(\mathbf{x}_t - \gamma_t \mathbf{g}_t)$$

- Entropic descent: $X = \Delta_d$, $\omega(\mathbf{x}) = \sum_{i=1}^d x_i \log x_i, V_\omega(\mathbf{x}, \mathbf{x}_t) = \mathrm{KL}(\mathbf{x} \mid \mathbf{x}_t)$.

$$\mathbf{x}_{t+1} \propto \mathbf{x}_t \odot \exp(-\gamma_t \mathbf{g}_t)$$

Here $\odot$ is element-wise multiplication.

Lemma 12.2

Let f be convex and $\omega(\cdot)$ be 1-strongly convex on X w.r.t. norm $\|\cdot\|$.

$$\gamma_t(f(\mathbf{x}_t) - f^*) \leq V_\omega(\mathbf{x}^*,\mathbf{x}_t) - V_\omega(\mathbf{x}^*,\mathbf{x}_{t+1}) + \frac{\gamma_t^2}{2}\|\mathbf{g}_t\|_*^2$$

Theorem 12.3

$$\min_{1\leq t\leq T} f(\mathbf{x}_t) - f^* \leq \frac{V_\omega(\mathbf{x}^*,\mathbf{x}_1) + \frac{1}{2}\sum_{t=1}^T \gamma_t^2 \|\mathbf{g}_t\|_*^2}{\sum_{t=1}^T \gamma_t}$$

Suppose f is B-Lipschitz continuous such that $|f(\mathbf{x}) - f(\mathbf{y})| \leq B\|\mathbf{x}-\mathbf{y}\|$, namely, $\|\mathbf{g}\|_* \leq B$ for any $\mathbf{g} \in \partial f(\mathbf{x})$. Define $R^2 := \sup_{\mathbf{x}\in X} V_\omega(\mathbf{x},\mathbf{x}_1)$, where $R \geq 0$ and set $\gamma_t = \frac{\sqrt{2}R}{B\sqrt{T}}$.

$$\min_{1\leq t\leq T} f(\mathbf{x}_t) - f^* \leq \mathcal{O}\left(\frac{BR}{\sqrt{T}}\right).$$

Mirror Descent for Smooth Objectives

Consider the problem $\min_X f(\mathbf{x})$, where X is closed and convex, and f is convex and L-smooth such that $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L\|\mathbf{x}-\mathbf{y}\|, \forall \mathbf{x},\mathbf{y} \in X$ for some norm $\|\cdot\|$. This further implies that

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

By setting $\gamma_t = 1/L$, the sequence of iterates $\{\frown_t\}$ generated by Mirror Descent:

$$\mathbf{x}_{t+1} = \operatorname*{argmin}_{\mathbf{x}\in X}\{V_\omega(\mathbf{x},\mathbf{x}_t) + \langle \gamma_t \nabla f(\mathbf{x}_t), \mathbf{x}\rangle\}$$

satisfies that

$$\min_{1\leq t\leq T} f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*) \leq \frac{L \cdot V_\omega(\mathbf{x}^*,\mathbf{x}_1)}{T}$$

(Proof see Hw12 Ex2)

Subgradient descent: special case with $\|\cdot\| = \|\cdot\|_2$ and $\omega(\cdot) = \frac{1}{2}\|\cdot\|_2^2$.

Optimization over simplex

Assume $\|\mathbf{g}\|_\infty \leq 1, \forall \mathbf{g} \in \partial f(\mathbf{x})$ and $X = \Delta_d$. Set $\mathbf{x}_1 = [1/d; \ldots; 1/d]$.

- Subgradient Descent: $\mathcal{O}\left(\frac{\sqrt{d}}{\sqrt{T}}\right)$, where $B = \mathcal{O}(\sqrt{d}), R = \mathcal{O}(1)$.
- Mirror Descent: $\mathcal{O}\left(\frac{\sqrt{\log d}}{\sqrt{T}}\right)$, where $B = \mathcal{O}(1), R = \mathcal{O}(\sqrt{\log d})$.

Part B: Smoothing Techniques

Convex Conjugate Theory

Conjugate Function

The conjugate function of f is

$$f^\star(\mathbf{y}) = \sup_{\mathbf{x}\in \mathrm{dom}(f)}\left\{\mathbf{y}^T\mathbf{x} - f(\mathbf{x})\right\}$$

also called Legendre-Fenchel transformation.

Fenchel's inequality

$f(\mathbf{x}) + f^*(\mathbf{y}) \geq \mathbf{x}^T\mathbf{y}, \forall \mathbf{x},\mathbf{y}$

Lemma 12.5

(1) Duality: If f is lower semi-continuous (l.s.c.) and convex, then $f^{\star\star} = f$.
Function f is l.s.c. if $f(\mathbf{x}) \leq \liminf_{t\to\infty} f(\mathbf{x}_t)$ for $\mathbf{x}_t \to \mathbf{x}$.

(2) Fenchel's inequality: $\mathbf{x}^T\mathbf{y} \leq f(\mathbf{x}) + f^\star(\mathbf{y})$.

(3) If f and g are l.s.c. and convex, then $(f+g)^\star(\mathbf{x}) = \inf_{\mathbf{y}}\left\{f^\star(\mathbf{y}) + g^\star(\mathbf{x}-\mathbf{y})\right\}$.

(4) If f is μ-strongly convex, then $f^\star$ is differentiable and $\frac{1}{\mu}$-smooth.

Examples

- Quadratic: $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x}$ where $\mathbf{Q} \succ 0, f^\star(y) = \frac{1}{2}\mathbf{y}^T\mathbf{Q}^{-1}\mathbf{y}$
- Negative entropy: $f(\mathbf{x}) = \sum_{i=1}^n x_i \log(x_i), f^\star(\mathbf{y}) = \sum_{i=1}^n e^{y_i - 1}$
- Negative logarithm: $f(\mathbf{x}) = -\sum_{i=1}^n \log(x_i), f^\star(\mathbf{y}) = -\sum_{i=1}^n \log(-y_i) - n$.
- Norm: $f(\mathbf{x}) = \|\mathbf{x}\|, f^\star(\mathbf{y}) = \begin{cases} 0, & \|\mathbf{y}\|_* \leq 1 \\ +\infty, & \|\mathbf{y}\|_* > 1 \end{cases}$

Part B: Smoothing Techniques

Nesterov's Smoothing for Convex Functions

Assume f is convex, then $f(\mathbf{x}) = (f^*)^* = \max_{\mathbf{y}}\{\mathbf{x}^\top\mathbf{y} - f^*(\mathbf{y}) - \mu \cdot d(\mathbf{y})\}$.
Nesterov's Smoothing is

$$f_\mu(\mathbf{x}) = \max_{\mathbf{y}\in \mathrm{dom}(f^\star)}\left\{\mathbf{x}^T\mathbf{y} - f^\star(\mathbf{y}) - \mu \cdot d(\mathbf{y})\right\} = \left(f^* + \mu d(\cdot)\right)^*$$

- Here $f^\star(\mathbf{y})$ is the convex conjugate of f.
- Proximity function: $d(\mathbf{y})$ is 1-strongly convex and nonnegative everywhere.
  - $d(\mathbf{y}) = \frac{1}{2}\|\mathbf{y}-\mathbf{y}_0\|_2^2$
  - $d(\mathbf{y}) = \frac{1}{2}\sum w_i\left(y_i - y_{0,i}\right)^2$ with $w_i \geq 1$;
  - $d(\mathbf{y}) = \omega(\mathbf{y}) - \omega(\mathbf{y}_0) - \nabla\omega(\mathbf{y}_0)^\top(\mathbf{y}-\mathbf{y}_0)$ with $\omega(\mathbf{x})$ being 1-strongly convex.
- Smoothness: Function $f_\mu(\mathbf{x})$ is $\frac{1}{\mu}$-smooth.
- Approximation: For convex f with bounded $\mathrm{dom}(f^\star)$, we have

$$f(\mathbf{x}) - \mu D^2 \leq f_\mu(\mathbf{x}) \leq f(\mathbf{x}), \text{ where } D^2 = \max_{\mathbf{y}\in\mathrm{dom}(f^\star)} d(\mathbf{y})$$

- Tradeoff between approximation error and optimization efficiency:

$$f(\mathbf{x}) - f^* \leq \underbrace{f(\mathbf{x}) - f_\mu(\mathbf{x})}_{\text{approximation error}} \underbrace{+ f_\mu(\mathbf{x}) - \min_x f_\mu(\mathbf{x})}_{\text{optimization error}}$$

- If we apply Accelerated Gradient Descent to solve the smoothed problem:

$$f(\mathbf{x}_t) - f^* \leq \mathcal{O}\left(\mu D^2 + \frac{R^2}{\mu t^2}\right) \leq \epsilon$$

To achieve accuracy $\epsilon > 0$, need $\mu = \mathcal{O}\left(\frac{\epsilon}{D^2}\right)$.
The number of AGD iterations is at most $T_\epsilon = \mathcal{O}\left(\frac{R}{\sqrt{\epsilon\mu}}\right) = \mathcal{O}\left(\frac{RD}{\epsilon}\right)$.

Moreau-Yosida Regularization for Convex Functions

$$f_\mu(\mathbf{x}) = \min_{\mathbf{y}}\left\{f(\mathbf{y}) + \frac{1}{2\mu}\|\mathbf{x}-\mathbf{y}\|_2^2\right\}$$

- Here $\mu > 0$ and $f_\mu(\mathbf{x})$ is called the Moreau envelope of $f(\mathbf{x})$.
- Huber function is Moreau envelope of $f(x) = |x|$:

$$f_\mu(x) = \begin{cases} \frac{x^2}{2\mu}, & |x| \leq \mu \\ |x| - \frac{\mu}{2}, & |x| > \mu \end{cases}$$

- M-Y Regularization is a special case of Nesterov's smoothing with $d(\mathbf{y}) = \frac{1}{2}\|\mathbf{y}\|^2$.

$$\begin{aligned} f_\mu(\mathbf{x}) &= \max_{\mathbf{y}}\left\{\mathbf{x}^T\mathbf{y} - f^\star(\mathbf{y}) - \frac{\mu}{2}\|\mathbf{y}\|_2^2\right\} \\ &= \left(f^\star + \frac{\mu}{2}\|\cdot\|_2^2\right)^\star(\mathbf{x}) \\ &= \inf_{\mathbf{y}}\left\{f(\mathbf{y}) + \frac{1}{2\mu}\|\mathbf{x}-\mathbf{y}\|_2^2\right\} \end{aligned}$$

- Smoothness: Function $f_\mu(\mathbf{x})$ is $\frac{1}{\mu}$-smooth.
- Exact Minimization: $\min_{\mathbf{x}} f(\mathbf{x}) = \min_{\mathbf{x}} f_\mu(\mathbf{x})$.

$$\text{prox}_{\mu f}(\mathbf{x}) := \operatorname*{argmin}_{\mathbf{y}} \left\{ f(\mathbf{y}) + \frac{1}{2\mu}\|\mathbf{x}-\mathbf{y}\|_2^2 \right\}$$

- Gradient of smooth function: (based on Danskin's theorem or Fenchel duality)

$$\nabla f_\mu(\mathbf{x}) = \frac{1}{\mu}\left(\mathbf{x} - \text{prox}_{\mu f}(\mathbf{x})\right)$$

- GD on smooth $f_\mu(\mathbf{x})$ reduces to proximal minimization on $f(\mathbf{x})$:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \mu \nabla f_\mu(\mathbf{x}_t) \Longleftrightarrow \mathbf{x}_{t+1} = \text{prox}_{\mu f}(\mathbf{x}_t)$$

Proximal Operators

The proximal operator of convex function g at $\mathbf{x}$ is defined as

$$\text{prox}_f(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{y}} \left\{ f(\mathbf{y}) + \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \right\}$$

For continuous convex function f, $\text{prox}_f(\mathbf{x})$ exists and is unique.
For many nonsmooth functions, proximal operators can be computed efficiently (closed form solution, low-cost computation, polynomial time).

Properties

Let g be a convex function with $\text{dom}(g) = \mathbb{R}^d$. Then we have

- (Subgradient characterization) $\mathbf{y} = \text{prox}_g(\mathbf{x}) \Longleftrightarrow \mathbf{x} - \mathbf{y} \in \partial g(\mathbf{y})$.
- (Fixed Point) A point $\mathbf{x}^*$ minimizes $g(\mathbf{x}) \Longleftrightarrow \mathbf{x}^* = \text{prox}_g(\mathbf{x}^*)$.
- (Non-expansiveness) $\left\|\text{prox}_g(\mathbf{x}) - \text{prox}_g(\mathbf{y})\right\|_2 \leq \|\mathbf{x}-\mathbf{y}\|_2$.
  (Proof see Hw12 Ex3)

Examples

If $f(\mathbf{x}) = \delta_X(\mathbf{x}) = \begin{cases} 0, & \mathbf{x} \in X \\ +\infty, & \mathbf{x} \notin X \end{cases}$, then $\text{prox}_f(\mathbf{x}) = \Pi_X(\mathbf{x})$ is the projection.
If $f(\mathbf{x}) = \mu\|\mathbf{x}\|_1$, then $\text{prox}_f(\mathbf{x})$ is the soft thresholding operator.

$$\text{prox}_{\mu|\cdot|}(x_i) = \begin{cases} x_i - \mu & \text{if } x_i > \mu \\ 0 & \text{if } |x_i| \leq \mu \\ x_i + \mu & \text{if } x_i < -\mu \end{cases}$$

Equivalently, $\text{prox}_{\mu\|\cdot\|_1}(\mathbf{x}) = \text{sign}(\mathbf{x}) \odot \max\{|\mathbf{x}| - \mu, 0\}$

Proximal Point Algorithm

$$\text{PPA:} \quad \mathbf{x}_{t+1} = \text{prox}_{\lambda_t f}(\mathbf{x}_t)$$

Theorem 12.7 Convergence of PPA

If f is convex, then for any $T \geq 1$,

$$f(\mathbf{x}_{T+1}) - f^* \leq \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|_2^2}{2\sum_{t=1}^T \lambda_t}.$$

Setting $\lambda_t = \lambda$, this implies a $\mathcal{O}(1/t)$ convergence rate.

Smoothing Techniques for Nonconvex Functions

Lasry-Lions Regularization

Handout12 Page 43

Randomized Smoothing

Handout12 Page 44

13. Min-Max Optimization

13.1 Min-Max Optimization

Min-Max Optimization

Let $\mathcal{X} \subset \mathbb{R}^d, \mathcal{Y} \subset \mathbb{R}^p$ and $\phi: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Consider the min-max problem:

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{x}, \mathbf{y})$$

Zero-sum Matrix Games

2-players games where players have opposite evaluations of outcomes:

- I (resp. J) non-empty finite set of strategies of player 1 (resp. player 2).
- payoff of player 1 given by a real-valued $I \times J$ matrix $\mathbf{A}$ (resp. $-\mathbf{A}$ for player 2).
- Set of mixed strategies $\Delta(I) = \left\{\mathbf{x} \in \mathbb{R}^{|I|} : \mathbf{x}_i \geq 0, i \in I, \sum_{i\in I} \mathbf{x}_i = 1\right\}$ of player 1 (resp. $\Delta(J)$ for player 2).

$$\min_{\mathbf{x}\in\Delta(I)} \max_{\mathbf{y}\in\Delta(J)} \mathbf{x}^T \mathbf{A}\mathbf{y}$$

Nonsmooth Optimization

Let f, g be convex nonsmooth functions, $\mathbf{A} \in \mathbb{R}^{p\times d}$ a matrix and consider the problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) + g(\mathbf{A}\mathbf{x})$$

Recall that $g(\mathbf{A}\mathbf{x}) = \max_{\mathbf{y}\in\mathbb{R}^p} \langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle - g^*(\mathbf{y})$ where g^* is the Fenchel conjugate.
Then the problem is equivalent to Min-Max reformulation:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \max_{\mathbf{y}\in\mathbb{R}^p} f(\mathbf{x}) + \langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle - g^*(\mathbf{y})$$

Examples: $g(\mathbf{z}) = \|\mathbf{z} - \mathbf{b}\|_1, g(\mathbf{z}) = \|\mathbf{z} - \mathbf{b}\|_2^2$ or $g(\mathbf{z}) = \iota_{\{\mathbf{b}\}}(\mathbf{z})$ (=0 if $\mathbf{z} = \mathbf{b}$, $+\infty$ otherwise) for which the Fenchel conjugate can be explicitly computed.

13.2 Saddle Points and Global Minimax Points

Saddle Points

Consider the min-max problem:

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{x}, \mathbf{y})$$

$(\mathbf{x}^*, \mathbf{y}^*)$ is a saddle point if $\phi(\mathbf{x}^*, \mathbf{y}) \leq \phi(\mathbf{x}^*, \mathbf{y}^*) \leq \phi(\mathbf{x}, \mathbf{y}^*)$ for any $\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}$

- Game interpretation: Nash equilibrium
- No player has the incentive to make unilateral change at NE.
- Simultaneous game

Global Minimax Points

$(\mathbf{x}^*, \mathbf{y}^*)$ is a global minimax point if $\phi(\mathbf{x}^*, \mathbf{y}) \leq \phi(\mathbf{x}^*, \mathbf{y}^*) \leq \max_{\mathbf{y}'\in\mathcal{Y}} \phi(\mathbf{x}, \mathbf{y}')$ for any $\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}$

- Game interpretation: Stackelberg equilibrium
- Best response to the best response.
- Sequential game

Primal and Dual Problems

$$(P): \quad \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{x}, \mathbf{y}) := \min_{\mathbf{x}\in\mathcal{X}} \overline{\phi}(\mathbf{x})$$

$$(D): \quad \max_{\mathbf{y}\in\mathcal{Y}} \min_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x}, \mathbf{y}) := \max_{\mathbf{y}\in\mathcal{Y}} \underline{\phi}(\mathbf{y})$$

Max-Min Inequality

$$\max_{\mathbf{y}\in\mathcal{Y}} \min_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x}, \mathbf{y}) \leq \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{x}, \mathbf{y})$$

Lemma 12.1 Characterization of Saddle Points

$(\mathbf{x}^*, \mathbf{y}^*)$ is a saddle point iff

$$\max_{\mathbf{y}\in\mathcal{Y}} \min_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{x}, \mathbf{y})$$

and $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} \overline{\phi}(\mathbf{x})$, $\mathbf{y}^* \in \operatorname{argmax}_{\mathbf{y}\in\mathcal{Y}} \underline{\phi}(\mathbf{y})$
Invoking the definition of saddle point, we have

$$\begin{aligned} \max_{\mathbf{y}\in\mathcal{Y}} \min_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x}, \mathbf{y}) &\geq \min_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x}, \mathbf{y}^*) \geq \phi(\mathbf{x}^*, \mathbf{y}^*) \\ &\geq \max_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{x}^*, \mathbf{y}) \geq \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{x}, \mathbf{y}) \end{aligned}$$

13.3 Convex-Concave Min-Max Optimization

Convex-concave function

A function $\phi(\mathbf{x}, \mathbf{y}): \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is convex-concave if

- $\phi(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{x} \in \mathcal{X}$ for every fixed $\mathbf{y} \in \mathcal{Y}$;
- $\phi(\mathbf{x}, \mathbf{y})$ is concave in $\mathbf{y} \in \mathcal{Y}$ for every fixed $\mathbf{x} \in \mathcal{X}$.

Strongly-convex-strongly-concave function

A function $\phi(\mathbf{x}, \mathbf{y}): \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is strongly-convex-strongly-concave if there exist constants $\mu_1, \mu_2 > 0$ such that

- $\phi(\mathbf{x}, \mathbf{y})$ is μ_1-strongly convex in $\mathbf{x} \in \mathcal{X}$ for every fixed $\mathbf{y} \in \mathcal{Y}$;
- $\phi(\mathbf{x}, \mathbf{y})$ is μ_2-strongly concave in $\mathbf{y} \in \mathcal{Y}$ for every fixed $\mathbf{x} \in \mathcal{X}$.

Thm 12.4 Minimax Theorem

If $\mathcal{X}$ and $\mathcal{Y}$ are closed convex sets and one of them is bounded, and $\phi(\mathbf{x}, \mathbf{y})$ is a continuous convex-concave function, then there exists a saddle point on $\mathcal{X} \times \mathcal{Y}$ and

$$\max_{\mathbf{y}\in\mathcal{Y}} \min_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{x}, \mathbf{y})$$

Here $\mathbf{x}, \mathbf{y}$ are arbitrary values, not necessarily a saddle point.

13.4 First-order Methods

Duality Gap: Accuracy Measure of Minimax Optimization

For convex-concave minimax optimization, saddle points exist.
We measure the optimality via the duality gap.

$$\text{duality gap} := \max_{\mathbf{y}\in\mathcal{Y}} \phi(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x}, \hat{\mathbf{y}}) \geq 0.$$

- When duality gap $= 0$, $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is a saddle point.
- When duality gap $\leq \epsilon$, $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is an ϵ-saddle point.

Gradient Descent Ascent (GDA)

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}\left(\mathbf{x}_t - \gamma \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t)\right)$$
$$\mathbf{y}_{t+1} = \Pi_{\mathcal{Y}}\left(\mathbf{y}_t + \gamma \nabla_{\mathbf{y}} \phi(\mathbf{x}_t, \mathbf{y}_t)\right)$$

Strongly-Convex-Strongly-Concave (SC-SC) Setting

- μ-strongly convex about $\mathbf{x}$ and strongly concave about $\mathbf{y}$:

$$\begin{aligned} \phi(\mathbf{x}_1, \mathbf{y}) &\geq \phi(\mathbf{x}_2, \mathbf{y}) + \nabla_{\mathbf{x}} \phi(\mathbf{x}_2, \mathbf{y})^\top (\mathbf{x}_1 - \mathbf{x}_2) \\ &\quad + \frac{\mu}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|^2 \\ -\phi(\mathbf{x}_1, \mathbf{y}_1) &\geq -\phi(\mathbf{x}, \mathbf{y}_2) - \nabla_{\mathbf{y}} \phi(\mathbf{x}, \mathbf{y}_2)^\top (\mathbf{y}_1 - \mathbf{y}_2) \\ &\quad + \frac{\mu}{2}\|\mathbf{y}_1 - \mathbf{y}_2\|^2 \end{aligned}$$

- l-Lipschitz smooth jointly in $\mathbf{x}$ and $\mathbf{y}$:

$$\begin{aligned} &\left\|\nabla_{\mathbf{x}} \phi(\mathbf{x}_1, \mathbf{y}_1) - \nabla_{\mathbf{x}} \phi(\mathbf{x}_2, \mathbf{y}_2)\right\| \\ &\leq L\left(\|\mathbf{x}_1 - \mathbf{x}_2\| + \|\mathbf{y}_1 - \mathbf{y}_2\|\right) \\ &\left\|\nabla_{\mathbf{y}} \phi(\mathbf{x}_1, \mathbf{y}_1) - \nabla_{\mathbf{y}} \phi(\mathbf{x}_2, \mathbf{y}_2)\right\| \\ &\leq L\left(\|\mathbf{x}_1 - \mathbf{x}_2\| + \|\mathbf{y}_1 - \mathbf{y}_2\|\right) \end{aligned}$$

- There exists a unique saddle point $(\mathbf{x}^*, \mathbf{y}^*)$

Thm 12.5 Convergence of GDA for SC-SC Setting

In SC-SC setting, GDA with stepsize $\eta < \frac{\mu}{2L^2}$ converges linearly:

$$\left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|^2 + \left\|\mathbf{y}_{t+1} - \mathbf{y}^*\right\|^2$$

When $\eta = \frac{\mu}{4L^2}$,

$$\begin{aligned}&\left\|\mathbf{x}_T-\mathbf{x}^*\right\|^2+\left\|\mathbf{y}_T-\mathbf{y}^*\right\|^2\\ \leq&\left(1-4\mu^2/L^2\right)^T\left(\left\|\mathbf{x}_0-\mathbf{x}^*\right\|^2+\left\|\mathbf{y}_0-\mathbf{y}^*\right\|^2\right)\end{aligned}$$

It implies a complexity of $\mathcal{O}\left(\kappa^2\log\frac{1}{\epsilon}\right)$ with $\kappa = L/\mu$ being condition number.

Extragradient (EG)

$$\begin{aligned}\mathbf{x}_{t+\frac{1}{2}}&=\Pi_{\mathcal{X}}\left(\mathbf{x}_t-\eta\nabla_{\mathbf{x}}\phi\left(\mathbf{x}_t,\mathbf{y}_t\right)\right)\\ \mathbf{y}_{t+\frac{1}{2}}&=\Pi_{\mathcal{Y}}\left(\mathbf{y}_t+\eta\nabla_{\mathbf{y}}\phi\left(\mathbf{x}_t,\mathbf{y}_t\right)\right)\\ \mathbf{x}_{t+1}&=\Pi_{\mathcal{X}}\left(\mathbf{x}_t-\eta\nabla_{\mathbf{x}}\phi\left(\mathbf{x}_{t+\frac{1}{2}},\mathbf{y}_{t+\frac{1}{2}}\right)\right)\\ \mathbf{y}_{t+1}&=\Pi_{\mathcal{Y}}\left(\mathbf{y}_t+\eta\nabla_{\mathbf{y}}\phi\left(\mathbf{x}_{t+\frac{1}{2}},\mathbf{y}_{t+\frac{1}{2}}\right)\right)\end{aligned}$$

Thm 12.6 EG for C-C Setting

Assume ϕ is convex-concave, L-Lipschitz smooth, $\mathcal{X}$ has diameter $D_{\mathcal{X}}$, and $\mathcal{Y}$ has diameter $D_{\mathcal{Y}}$, then EG with stepsize $\eta \leq \frac{1}{2L}$ satisfies

$$\max_{\mathbf{y}\in\mathcal{Y}}\phi\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_{t+\frac{1}{2}},\mathbf{y}\right)-\min_{\mathbf{x}\in\mathcal{X}}\phi\left(\mathbf{x},\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_{t+\frac{1}{2}}\right)\leq\frac{D_{\mathcal{X}}^2+D_{\mathcal{Y}}^2}{2\eta T}$$

$\mathcal{O}(1/T)$ convergence rate for averaged iterates at "mid-point".
$\mathcal{O}(1/T)$ rate is optimal.

Thm 12.7

In SC-SC setting, EG with stepsize $\eta = \frac{1}{8L}$ converges linearly:

$$\begin{aligned}&\left\|\mathbf{x}_{t+1}-\mathbf{x}^*\right\|^2+\left\|\mathbf{y}_{t+1}-\mathbf{y}^*\right\|^2\\ \leq&\left(1-\frac{\mu}{4L}\right)\left\{\left\|\mathbf{x}_t-\mathbf{x}^*\right\|^2+\left\|\mathbf{y}_t-\mathbf{y}^*\right\|^2\right\}\end{aligned}$$

This $\mathcal{O}\left(\kappa\log\frac{1}{\epsilon}\right)$ complexity is optimal for SC-SC setting.

Optimistic GDA

$$\begin{aligned}\mathbf{x}_{t+\frac{1}{2}}&=\Pi_{\mathcal{X}}\left(\mathbf{x}_t-\eta\nabla_{\mathbf{x}}\phi\left(\mathbf{x}_{t-\frac{1}{2}},\mathbf{y}_{t-\frac{1}{2}}\right)\right)\\ \mathbf{y}_{t+\frac{1}{2}}&=\Pi_{\mathcal{Y}}\left(\mathbf{y}_t+\eta\nabla_{\mathbf{y}}\phi\left(\mathbf{x}_{t-\frac{1}{2}},\mathbf{y}_{t-\frac{1}{2}}\right)\right)\\ \mathbf{x}_{t+1}&=\Pi_{\mathcal{X}}\left(\mathbf{x}_t-\eta\nabla_{\mathbf{x}}\phi\left(\mathbf{x}_{t+\frac{1}{2}},\mathbf{y}_{t+\frac{1}{2}}\right)\right)\\ \mathbf{y}_{t+1}&=\Pi_{\mathcal{Y}}\left(\mathbf{y}_t+\eta\nabla_{\mathbf{y}}\phi\left(\mathbf{x}_{t+\frac{1}{2}},\mathbf{y}_{t+\frac{1}{2}}\right)\right)\end{aligned}$$

Equivalent formulation:

$$\begin{cases}\mathbf{x}_{t+1}=\mathbf{x}_t-2\eta\nabla_{\mathbf{x}}\phi\left(\mathbf{x}_t,\mathbf{y}_t\right)+\eta\nabla_{\mathbf{x}}\phi\left(\mathbf{x}_{t-1},\mathbf{y}_{t-1}\right)\\ \mathbf{y}_{t+1}=\mathbf{y}_t-2\eta\nabla_{\mathbf{y}}\phi\left(\mathbf{x}_t,\mathbf{y}_t\right)+\eta\nabla_{\mathbf{y}}\phi\left(\mathbf{x}_{t-1},\mathbf{y}_{t-1}\right)\end{cases}$$

Proximal Point Algorithm (PPA)

$$\begin{aligned}&\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)\\ \leftarrow&\operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}}\operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}}\left\{\phi(\mathbf{x},\mathbf{y})+\frac{1}{2\eta}\|\mathbf{x}-\mathbf{x}_t\|^2-\frac{1}{2\eta}\|\mathbf{y}-\mathbf{y}_t\|^2\right\}\end{aligned}$$

PPA has been shown to converge with $\mathcal{O}(1/T)$ rate in convex-concave case.

Implicit Update of PPA

$$\begin{aligned}\mathbf{x}_{t+1}&=\Pi_{\mathcal{X}}\left(\mathbf{x}_t-\eta\nabla_{\mathbf{x}}\phi\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)\right)\\ \mathbf{y}_{t+1}&=\Pi_{\mathcal{Y}}\left(\mathbf{y}_t+\eta\nabla_{\mathbf{y}}\phi\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)\right)\end{aligned}$$

Connections between PPA, EG and OGDA

Handout13 Page 33

13.5 Concave Games, Variational Inequalities

Variational Inequality Problem (VI)

Let $\mathcal{Z}\subset\mathbb{R}^d$ be a nonempty subset and consider a mapping $F:\mathcal{Z}\to\mathbb{R}^d$.

VI Problem: Find $\mathbf{z}^*\in\mathcal{Z}$ such that $\langle F(\mathbf{z}^*),\mathbf{z}-\mathbf{z}^*\rangle\geq 0$ for all $\mathbf{z}\in\mathcal{Z}$.

Existence: If $\mathcal{Z}$ is a nonempty convex compact subset of $\mathbb{R}^d$ and $F:\mathcal{Z}\to\mathbb{R}^d$ is continuous, then there exists a solution z^* to (VI).

Variational Inequalities with Monotone Operators

The operator $F:\mathcal{Z}\to\mathbb{R}^d$ is:

- monotone if

$$\langle F(\mathbf{u})-F(\mathbf{v}),\mathbf{u}-\mathbf{v}\rangle\geq 0\quad\forall\mathbf{u},\mathbf{v}\in\mathcal{Z}$$

- μ-strongly-monotone ($\mu>0$) if

$$\langle F(\mathbf{u})-F(\mathbf{v}),\mathbf{u}-\mathbf{v}\rangle\geq\mu\|\mathbf{u}-\mathbf{v}\|^2\quad\forall\mathbf{u},\mathbf{v}\in\mathcal{Z}$$

Weak Solution of VI

- (Strong) solution (of Stampacchia VI): find $\mathbf{z}^*\in\mathcal{Z}$ such that:

$$\left\langle F\left(\mathbf{z}^*\right),\mathbf{z}-\mathbf{z}^*\right\rangle\geq 0\forall\mathbf{z}\in\mathcal{Z}.$$

- Weak solution (of Minty VI): find $\mathbf{z}^*\in\mathcal{Z}$ such that:

$$\left\langle F(\mathbf{z}),\mathbf{z}-\mathbf{z}^*\right\rangle\geq 0\forall\mathbf{z}\in\mathcal{Z}.$$

- If F is monotone, then a strong solution is also a weak solution.
- If F is continuous, then a weak solution is also a strong solution.
- We use $\epsilon_{\mathrm{VI}}(\hat{\mathbf{z}}):=\max_{\mathbf{u}\in\mathcal{Z}}\langle F(\mathbf{u}),\mathbf{u}-\hat{\mathbf{z}}\rangle$ to measure the inaccuracy of a solution $\hat{\mathbf{z}}$.