Convex Optimization Notes

Wilson Pan

February 12, 2026

Contents

1   Jan 13

Definition 1.1. *An **optimization problem** is defined as*

$$\min_{x \in \mathcal{C}} f_0(x) \quad where \quad \forall i \in [m], \ f_i(x) \le b_i$$

where $\mathcal{C}$ is the constraint set and $f_0 \colon \mathcal{C} \to \mathbb{R}$.

Definition 1.2. *The **standard form** of an optimization problem is*

$$\min_{x \in \mathcal{D}} f_0(x) \quad subject\ to \quad \forall i \in [m], \ f_i(x) \le 0; \quad \forall j \in [p], \ h_j(x) = 0$$

where $x \in \mathrm{dom}(f_i) \cap \mathrm{dom}(h_j) =: \mathcal{D}$.
f_0 is termed the objective function, *and x is the optimization variable.*
f_i and h_j are constraint functions.

Definition 1.3. *A **feasible solution** is any $x \in \mathcal{D}$ that satisfies all the constraints.*

Definition 1.4. *The **optimal solution** is $x^* \in \arg\min_{x \in \mathcal{D}} f_0(x)$.*

Remark 1.5. *If no feasible solution exists, we define $\min_{x \in \mathcal{C}} f(x) = +\infty$ and, similarly, $\max_{x \in \mathcal{C}} f(x) = -\infty$.*

2 Jan 15

Definition 2.1. *A set C is **convex** if $\forall x_1, x_2 \in C$ and $\forall \theta \in [0,1]$, we have $\theta x_1 + (1-\theta)x_2 \in C$.*

Definition 2.2. *A function $f : \mathcal{X} \to \mathbb{R}$ is **convex** if for all $x_1, x_2 \in \mathcal{X}$ and $\theta \in [0,1]$,*

$$f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta)f(x_2) \tag{*}$$

Remark 2.3. *Convex functions are differentiable almost everywhere.*

Theorem 2.4. *If f is both convex and concave, then f is affine (i.e., linear plus constant).*

Definition 2.5. *f is **strictly convex** if $(*)$ holds with strict inequality $<$ whenever $x_1 \neq x_2$ and $\theta \in (0,1)$.*

Definition 2.6. *A **convex optimization problem** is one where f_0, f_i, and $\mathcal{C}$ are convex, and h_j are affine/linear. That is,*

$$\min_{x \in \mathcal{C}} f_0(x) \quad \text{such that} \quad f_i(x) \leq 0 \;\; \forall i \in [m], \quad h_j(x) = 0 \;\; \forall j \in [p]$$

Theorem 2.7. *The set of feasible solutions in Definition 2.6 is convex.*

Theorem 2.8. *x^* is a **local minimizer** if there exists $\epsilon > 0$ such that for all y with $\|x^* - y\| \leq \epsilon$,*

$$f(x^*) \leq f(y)$$

Theorem 2.9. *x^* is a **local maximizer** if there exists $\epsilon > 0$ such that for all y with $\|x^* - y\| \leq \epsilon$,*

$$f(x^*) \geq f(y)$$

Theorem 2.10. *For any convex optimization problem, every local minimum is a global minimum.*

Proof. Suppose $\hat{x}$ is a local minimizer not equal to global minimizer x^*. Take ϵ as any witness to $\hat{x}$ being a local minimum. Let

$$y = \frac{\epsilon}{\|\hat{x} - x^*\|}x^* + \left(1 - \frac{\epsilon}{\|\hat{x} - x^*\|}\right)\hat{x}$$

Note: $\|\hat{x} - x^*\| \leq \epsilon$, otherwise $\hat{x}$ is not a local minimizer in that neighborhood.

$$y - \hat{x} = \frac{\epsilon(x^* - \hat{x})}{\|x^* - \hat{x}\|} \quad ; \quad \|y - \hat{x}\| = \epsilon$$

Since f is convex,

$$f(y) \leq \frac{\epsilon}{\|\hat{x} - x^*\|}f(x^*) + \left(1 - \frac{\epsilon}{\|\hat{x} - x^*\|}\right)f(\hat{x}) < f(\hat{x})$$

Thus, a contradiction. $\qquad\square$

3 Jan 20

Definition 3.1. *A **convex combination** of x_1, x_2 is $\theta_1 x_1 + \theta_2 x_2$ where $\theta_1, \theta_2 \geq 0$ and $\theta_1 + \theta_2 = 1$.*

Definition 3.2. *An **affine combination** of x_1, x_2 is $\theta_1 x_1 + \theta_2 x_2$ where $\theta_1 + \theta_2 = 1$.*

Definition 3.3. *A **linear combination** of x_1, x_2 is $\theta_1 x_1 + \theta_2 x_2$ where $\theta_1, \theta_2 \in \mathbb{R}$.*

Definition 3.4. *A **conic combination** of x_1, x_2 is $\theta_1 x_1 + \theta_2 x_2$ where $\theta_1, \theta_2 \geq 0$.*

Definition 3.5. *Given a set C, the **convex hull** of C is*

$$\text{conv}(C) = \left\{ \sum_{i=1}^{k} \theta_i x_i \;\middle|\; x_i \in C, \, \theta_i \in [0,1], \, \sum_{i=1}^{k} \theta_i = 1 \right\}$$

Remark 3.6. *The following are true:*

1. *$C \subseteq \text{conv}(C)$*

2. *$\text{conv}(C)$ is convex*

3. *It is the smallest convex set containing C*

4. *If a convex set $S \supseteq C$ then $S \supseteq \text{conv}(C)$*

Theorem 3.7. *Any closed convex set can be written as $\overline{\text{conv}}(C)$ for some set C.*

Definition 3.8. *The **relative interior** of C is defined as*

$$\text{relint}(C) = \{ x \in C : \exists \epsilon > 0, \; B(x, \epsilon) \cap \text{Aff}(C) \subseteq C \}$$

Definition 3.9. *C is a **cone** if $\alpha x \in C$ whenever $x \in C$ and $\alpha \geq 0$.*

Definition 3.10. *Given a set C, the **conic hull** of C is*

$$\text{conic}(C) = \left\{ \sum_{i=1}^{k} \theta_i x_i : x_i \in C, \; \theta_i \geq 0 \right\}$$

Theorem 3.11. *The conic hull of C is the smallest convex cone containing C.*

Definition 3.12. *The ℓ_p **norm** is defined as*

$$\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p}$$

Remark 3.13. *The following are true:*

1. *For $p \in (0, 1)$: $\|x\|_p$ is not a convex function*

2. *For $p \geq 1$: $\|x\|_p$ is convex*

3. *For $p > 1$: $\|x\|_p$ is strictly convex*

Example 3.14. *Examples of convex sets:*

1. *Hyperplane: $\{x : a^T x = b\}$*

2. *Halfspace: $\{x : a^T x \leq b\}$*

3. *Polyhedron: $\{x \in \mathbb{R}^d : Ax \leq b, \, Cx = d\}$*

4. *Polytope: a bounded polyhedron.*

Theorem 3.15. *A set S is **strictly convex** if for all $x_1 \neq x_2$ and $\theta \in (0, 1)$, $\theta x_1 + (1 - \theta) x_2 \in \text{int}(S)$.*

Definition 3.16. *The **normal cone** is defined to be*

$$N_C(x) = \{ g : g^T (y - x) \leq 0, \; \forall y \in C \}$$

Remark 3.17. *If $x \in \text{int}(C)$ then $N_C(x) = \{0\}$.*

Theorem 3.18. *If f is differentiable, then x^* is optimal if and only if $-\nabla f(x^*) \in N_C(x^*)$.*

4   Jan 22

Theorem 4.1. *If f is convex and differentiable, and C is a convex set, then any optimal solution x^* to $\min_{x \in C} f(x)$ must satisfy $-\nabla f(x^*) \in N_C(x^*)$.*

Theorem 4.2. *The set of optimal solutions to a convex optimization problem is a convex set.*

Definition 4.3. *If C is convex then*

1. **Translation:** $C + a = \{x : x - a \in C\}$

2. **Scaling:** $\alpha C = \{x : \frac{x}{\alpha} \in C\}$

3. **Intersection:** *If $\{C_\alpha\}_{\alpha \in A}$ is a collection of convex sets, then $\bigcap_{\alpha \in A} C_\alpha$ is convex.*

Theorem 4.4. *The following are true:*

1. *If $C \subseteq \mathbb{R}^n$ is convex, $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, then $AC + b = \{Ax + b : x \in C\}$ is a convex set.*

2. *If $f(x) = Ax + b$, then $f^{-1}(C)$ is convex.*

3. *$C_1 + C_2 = \{x + y : x \in C_1, y \in C_2\}$ is convex.*

4. *If $C_1 \subseteq \mathbb{R}^m$, $C_2 \subseteq \mathbb{R}^n$ then $C_1 \times C_2 = \{(x, y) \in \mathbb{R}^{m+n} : x \in C_1, y \in C_2\}$ is convex.*

5. *For any $C \subseteq \mathbb{R}^n \times \mathbb{R}^m_{>0}$, define*

$$P(C) = \left\{ \left(\frac{x_1}{t}, \ldots, \frac{x_n}{t} \right) : (x_1, \ldots, x_n, t) \in C \right\}$$

   If C is convex, so are $P(C)$ and $P^{-1}(C)$.

Definition 4.5.

$$\text{epi}(f) = \{(x, t) : x \in \text{dom}(f), t \geq f(x)\}$$

Definition 4.6. *First Order Definition of Convexity: If f is differentiable, then f is convex if and only if for all $x, y \in \text{dom}(f)$,*

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

Figure 1: First Order Definition of Convexity

Second Order Definition of Convexity: If f is twice differentiable, then f is convex if and only if $\nabla^2 f(x) \succeq 0$.

Definition 4.7. *We say $A \succeq B$ if $A - B$ is positive semi-definite. This is equivalent to*

$$a^T \nabla^2 f(x) a \geq 0 \quad \text{for all } a \in \mathbb{R}^d$$

Definition 4.8. *The* **subdifferential** *is*

$$\partial f(x) = \left\{ g : f(y) \geq f(x) + g^T(y - x), \ \forall y \in C \right\}$$

Any such $g \in \partial f(x)$ is called a **subgradient.**
If f is differentiable at x then $\partial f(x) = \{\nabla f(x)\}$.

Theorem 4.9. *f is convex if and only if ∂f is non-empty.*

5   Jan 27

Definition 5.1. *Gradient Monotonicity: If f is differentiable, then f is convex if and only if $\nabla f(x)$ is monotone. So we can conclude*

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq 0 \iff f \text{ is convex.}$$

Theorem 5.2. *For a $A \in \mathbb{R}^{m \times n}$ we can decompose*

$$A = U \sum V^t.$$

where $U \in \mathbb{R}^{m \times k}$ with orthonormal column, $\sigma \in \mathbb{R}^{k \times k}$ is a diagonal matrix with non-negative entries, and $V \in \mathbb{R}^{n \times k}$ with orthonormal column.

Definition 5.3. *(u_i, v_i, σ_i) form a s.v. triplet if $Av_i = \sigma_i u_i$ and $A^T u_i = \sigma_i v_i$ $||u_i|| = ||v_i|| = 1$ and $\sigma_i \geq 0$*

Theorem 5.4. *For any $A \in \mathbb{R}^{n \times n}$ then $A^T A$ is always positive semi definite*

Proof. We can write $A = V \sum U$ then $A^T A = V \sum U^T U \sum V^T = V \sum^2 V^T$ □

Definition 5.5. *Spectral Radius: $\max_i \{ |\lambda_i| : \lambda_i \in \Lambda(A) \} = \rho(A)$*

Definition 5.6. *The norm must satisfy the following properties:*

1. *$||A|| \geq 0$*
2. *$||\alpha A|| = |\alpha| ||A||$*
3. *$||A|| = 0$ if and only if $A = 0$*
4. *$||A + A'|| \leq ||A|| + ||A'||$*

Definition 5.7. *Operator/Spectral Norm:*

$$||A||_{op} = ||A||_2 = \max_{||x||=1} ||Ax||_2.$$

Definition 5.8. *Frobenius Norm:*

$$||A||_F = \sqrt{\sum_{i,j} A_{ij}^2}.$$

Definition 5.9. *f is L-Lipschitz if*

$$|f(x) - f(y)| \leq L ||x - y||_2 \quad \forall x, y \in \text{dom}(f).$$

If f is differentiable then f is L-Lipschitz if and only if $|\nabla f| \leq L$

Definition 5.10. *Differentiable f is β-smooth if*

$$||\nabla f(x) - \nabla f(y)||_2 \leq \beta ||x - y||_2 \quad \forall x, y \in \text{dom}(f).$$

If f is twice differentiable then f is β-smooth if and only if $\nabla^2 f(x) \preceq \beta I$.

Theorem 5.11. *If f is twice differentiable and β-smooth then*

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{\beta}{2} ||y - x||_2^2 \quad \forall x, y \in \text{dom}(f).$$

6 Feb 3

Definition 6.1. *"Descent Direction"* any h for which $f(x + \eta h) \leq f(x)$.

Theorem 6.2. *Assume f differentiable and β-smooth with $x_{t+1} \leftarrow x_t - \eta \nabla f(x_t)$.*

1. *If $\eta \leq \frac{2}{\beta}$ then $f(x_{t+1}) \leq f(x_t)$*

2. *If $\eta \leq \frac{1}{\beta}$ then $f(x_{t+1}) \leq f(x_t) - \frac{\eta}{2}||\nabla f(x_t)||_2^2$*

Proof. Recall
$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{\beta}{2}||y - x||_2^2.$$
and
$$||\nabla f(y) - \nabla f(x)||_2 \leq \beta||y - x||_2.$$
Let $y \leftarrow x + h$ then $f(x + h) \leq f(x) + \nabla f(x)^T h + \frac{\beta}{2}||h||_2^2$

1. If $h = -\frac{2}{\beta}\nabla f(x)$ then $\nabla f(x)^T h = -\frac{2}{\beta}||\nabla f(x)||_2^2$ and $\frac{\beta}{2}||h||_2^2 = \frac{2}{\beta}||\nabla f(x)||_2^2$ so $f(x + h) \leq f(x)$

2. If $h = -\frac{1}{\beta}\nabla f(x)$ then $\nabla f(x)^T h = -\frac{1}{\beta}||\nabla f(x)||_2^2$ and $\frac{\beta}{2}||h||_2^2 = \frac{1}{2\beta}||\nabla f(x)||_2^2$.

$\square$

Theorem 6.3. *If f is differentiable, β smooth, and x^* be any optimizer. For any $k > 0$*
$$\min_{i=0,...,k} ||\nabla(f(x_i))||_2 \leq \sqrt{\frac{2\beta}{k}\left(\underbrace{f(x_0) - f(x^*)}_{\Delta_0}\right)}.$$

Proof. AFSOC $\forall i \in \{0, ..., k\}, ||\nabla f(x_i)|| > \sqrt{\frac{2\beta}{k}\Delta_0}$ Then we have
$$f(x_{i+1}) < f(x_i) - \frac{1}{2\beta}\left(\frac{2\beta\Delta_0}{k}\right). \qquad\qquad (\forall i \in \{0, ..., k\})$$
Continuing down the chain we have
$$f(x_{k+1}) < f(x_0) - \Delta_0 < f(x^*).$$
Thus a contradiction. $\square$

Theorem 6.4. *If f is β-smooth and convex, $\eta = \frac{1}{\beta}$ and x^* any minimum then*
$$f(x_k) - f(x^*) \leq \frac{\beta||x_0 - x^*||_2^2}{2k}.$$

Proof. Observe that
$$||x_{t+1} - x^*||^2 = ||x_t - x^* - \eta\nabla f(x_t)||^2 = ||x_t - x^*||^2 - \eta^2||\nabla f(x_t)||^2 - 2\eta\nabla f(x_t)^T(x_t - x^*).$$
Rearranging and by convexity we have
$$\begin{aligned}
f(x_t) - f(x^*) \leq \nabla f(x_t)^T(x_t - x^*) &= \frac{1}{2\eta}\left[\underbrace{||x_t - x^*||^2}_{\delta_t} - \underbrace{||x_{t+1} - x^*||^2}_{\delta_{t+1}} + \frac{\eta}{2}||\nabla f(x_t)||^2\right] \\
&\leq \frac{1}{2\eta}\left[\delta_t - \delta_{t+1} + 2\eta(f(x_t) - f(x_{t+1}))\right] \\
&= \frac{1}{2\eta}(\delta_t - \delta_{t+1}) + f(x_t) - f(x_{t+1})
\end{aligned}$$
So we have
$$f(x_{t+1}) - f(x^*) \leq \frac{\beta}{2}(\delta_t - \delta_{t+1}).$$

Adding across all terms

$$\sum_{i=0}^{k-1} f(x_{i+1}) - f(x^*) \leq \frac{\beta}{2} \left(||x_0 - x^*||^2 \right).$$

So we have the relation

$$k(f(x_k) - f(x^*)) \leq \sum_{i=0}^{k-1} f(x_{i+1}) - f(x^*) \leq \frac{\beta}{2} ||x_0 - x^*||^2.$$

$\square$

Definition 6.5. f *is α-strongly convex if*

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \frac{\alpha}{2}t(1-t)||x-y||_2^2.$$

Additionally, if f is differentiable then

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\alpha}{2}||y - x||_2^2.$$

If f is twice differentiable then f is α-strongly convex if and only if $\nabla^2 f(x) \succeq \alpha I$.

Theorem 6.6. *Assume f is β-smooth, α-strongly convex, $\eta = \frac{1}{\beta}$ then $||x_k - x^*||^2 \leq \left(1 - \frac{1}{\gamma} \right)^k ||x_0 - x^*||^2$. We define the conditional number as $\gamma = \frac{\beta}{\alpha}$*

Proof.

$$\frac{\alpha}{2}||x_k - x^*||^2 \leq \frac{\beta}{2} \left(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2 \right) \leq \left(1 - \frac{\alpha}{\beta} \right) ||x_t - x^*||^2 \leq (1 - \frac{1}{\gamma})||x_k - x^*||^2.$$

$\square$

7 Feb 5

Recall 7.1. *From previous class*

1. $f(x + \eta h) \approx f(x) + \eta h^T \nabla f(x)$ *and when we set* $h = -\nabla f(x)$ *we have*

$$f(x - \eta \nabla f(x)) \approx f(x) - \eta ||\nabla f(x)||_2^2.$$

   Backtracking Line Search: *Pick* $\gamma_1 \in (0, 1)$ *and* $\gamma_2 \in (0, 1)$, *start with initial step size* η. *While* $f(x - \eta \nabla f(x)) > f(x) - \gamma_1 \eta ||\nabla f(x)||_2^2$ *(Armijo condition not satisfied), set* $\eta \leftarrow \gamma_2 \eta$ *and retry. Once the condition is satisfied, use step size* η *to update* $x \leftarrow x - \eta \nabla f(x)$.

Recall 7.2. *g is a subgradient of f at x if*

$$f(y) \geq f(xlabel) + g^T(y - x) \quad \forall y \in \text{dom}(f).$$

Lemma 7.3. *The following are true:*

1. *$\partial f(x)$ makes sense for non convex functions too, could be empty*

2. *If f is convex, then for $x \in \text{RelInt}(\text{dom}(f))$ we have that $\partial f(x)$ is non-empty*

3. *$\partial f(x)$ is a convex*

4. *If f is convex and differentiable at x then $\partial f(x) = \{\nabla f(x)\}$*

5. *If $\partial f(x)$ is non-empty everywhere, f is convex.*

Theorem 7.4. *The following are true:*

1. *$\partial(af) = a\partial f(x)$*

2. *$\partial(f + g) = \partial f(x) + \partial g(x)$*

3. *If $g(x) = f(Ax + b)$ then $\partial g(x) = A^T \partial f(Ax + b)$*

Example 7.5. *For $f(x) = \max_{i=1,\dots,n} f_i(x)$ we have $\partial f(x) = \text{Conv}\left(\bigcup_{i=1,\dots,n} \partial f_i(x)\right)$*

Example 7.6. *For $f(x) = |x|$ we have $\partial f(x) = \text{sign}(x) = \max -x, x$ so $\partial f(0) = [-1, 1]$.*

Example 7.7. *Let C be a convex set and $I_C(x) = \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases}$ is convex Additionally, for $x \in C$,* $\partial I_C(x) = N_C(x) = \{g : g^T(y - x) \leq 0 \forall y \in C\}$

Proof. For $x \in C$, $I_C(y) \geq I_C(x) + g^T(y - x)$ for all $y \in C$ then we have

$$0 \geq 0 + g^T(y - x).$$

$\square$

Definition 7.8. *Subgradient Method: $x_{t+1} \leftarrow x_t - \eta g_x$ for some $g_x \in \partial f(x)$*

1. *Subgradients are in general not descent directions*

2. *The min norm subgradient is a descent direction*

Example 7.9. *$f(x) = ||x||_2 = \sqrt{\sum_{i=1}^n x_i^2}$ is convex, then for $x \neq 0$ we have $\partial f(x) = \frac{x}{||x||_2}$.*

$$\partial f(0) = \{g : ||g||_2 \leq 1\}.$$

8 Feb 10

Definition 8.1. *For the subgradient method for $\min_{x \in \mathbb{R}^n} f(x)$ and f convex where*

$$x_{t+1} = x_t - \eta_t g_t \text{ where } g_t \in \partial f(x_t).$$

We define the best iterate as

$$x_T^{(\text{best})} = \arg \min_{i=0,\dots,T} f(x_i).$$

Theorem 8.2. *Assume f is G-lipschitz and convex. Let $\|x_0 - x^*\| \le R$ then pick*

$$\eta_t = \frac{R}{G\sqrt{T}} \text{ guarantees that } f\left(x_T^{(\text{best})}\right) - f(x^*) \le \frac{GR}{\sqrt{T}}.$$

Theorem 8.3. *A convex function f is G-lipschitz iff*

$$\|g_x\| \le G \ \forall x \in \text{dom}(f) \text{ and } \forall g_x \in \partial f(x).$$

Theorem 8.4. *For nonconvex, differentiable f, f being G-lipschitz iff*

$$\|\nabla f(x)\| \le G \ \forall x \in \text{dom}(f).$$

Theorem 8.5. *Assume f is convex and G-Lipschitz, and that an optimal solution x^* exists with $\|x_0 - x^*\| \le R$ for some $R > 0$.*
Pick $\eta_t \to 0$ such that $\sum_{t=1}^{\infty} \eta_t = \infty$ and $\sum_{t=1}^{T} \eta_t^2 < \infty$ then

$$f\left(x_T^{(\text{best})}\right) \to f(x^*) \text{ as } T \to \infty.$$

Theorem 8.6. *For the subgradient method with step sizes $\{\eta_t\}_{t=1}^{T}$ on a convex, G-Lipschitz function f, we have*

$$f\left(x_T^{(\text{best})}\right) - f(x^*) \ \le \ \frac{\|x_0 - x^*\|^2 + G^2 \sum_{t=1}^{T} \eta_t^2}{2 \sum_{t=1}^{T} \eta_t}.$$

Verification of Theorem 8.2. Set $\eta_t = \frac{R}{G\sqrt{T}}$ for all t. With $\|x_0 - x^*\| \le R$:

$$\sum_{t=1}^{T} \eta_t = T \cdot \frac{R}{G\sqrt{T}} = \frac{R\sqrt{T}}{G},$$

$$\sum_{t=1}^{T} \eta_t^2 = T \cdot \frac{R^2}{G^2 T} = \frac{R^2}{G^2}.$$

Substituting into the bound (using $\|x_0 - x^*\|^2 \le R^2$):

$$f\left(x_T^{(\text{best})}\right) - f(x^*) \le \frac{R^2 + G^2 \cdot \frac{R^2}{G^2}}{2 \cdot \frac{R\sqrt{T}}{G}} = \frac{2R^2}{2R\sqrt{T}/G} = \frac{2R^2 \cdot G}{2R\sqrt{T}} = \frac{GR}{\sqrt{T}}.$$

Hence Theorem 8.2 holds with the $O(1/\sqrt{T})$ rate. $\qquad\qquad\square$

Proof.

$$\begin{aligned}
\|x_{t+1} - x^*\|^2 &= \|x_t - \eta_t g_t - x^*\|^2 \\
&= \|x_t - x^*\|^2 - 2\eta_t g_t^T (x_t - x^*) + \eta_t^2 \|g_t\|^2 \\
&\le \|x_t - x^*\|^2 + 2\eta_t (f(x^*) - f(x_t)) + \eta_t^2 G^2 \qquad (*)
\end{aligned}$$

Last step follows from $f(x^*) \ge f(x_t) + g_t^T(x^* - x_t) \iff -g_t^T(x_t - x^*) \ge f(x^*) - f(x_t)$.
Adding $(*)$ up from $0, \dots, T-1$ we have

$$\|x_T - x^*\|^2 \le \|x_0 - x^*\|^2 + G^2 \sum_{t=0}^{T-1} \eta_t^2 + 2 \sum_{t=0}^{T-1} \eta_t \left(f(x^*) - f(x_t)\right).$$

Figure 2: Subgradient method: the path may zigzag (subgradients need not be descent directions). The **best iterate** $x_T^{(\text{best})}$ minimizes f over $\{x_0, \ldots, x_T\}$. The bound shrinks when $\sum \eta_t$ is large (more "progress") and when $\|x_0 - x^*\|$ and $\sum \eta_t^2$ are small.

So

$$2 \sum_{t=0}^{T-1} \eta_t \left(f(x_T^{(\text{best})}) - f(x^*) \right) \leq 2 \sum_{t=0}^{T-1} \eta_t \left(f(x^*) - f(x_t) \right) \leq \|x_0 - x^*\|^2 + G^2 \sum_{t=0}^{T-1} \eta_t^2.$$

$\square$

Theorem 8.7. *Polyak's Stepsize: If $f(x^*)$ is known,*

$$\eta_t = \frac{f(x_t) - f(x^*)}{\|g_t\|^2}.$$

Theorem 8.8. *Given $C_1, \ldots, C_k$ convex sets find $x^* \in \bigcap_{i=1}^{k} C_i$.*

Proof. Define $f_i(x) = \min_{y \in C_i} \|x - y\|^2 = \text{dist}(x, C_i)$ and $f(x) = \max_{i=1,\ldots,k} f_i(x)$.
If $x^* \in C_1 \cap \ldots \cap C_k$ then $f(x^*) = 0$.
Recall: $\partial f(x) = \text{Conv} \left(\bigcup_{i=1,\ldots,k} \partial f_i(x) \right)$.
Let $P_C(x) = \arg\min_{y \in C} \|x - y\|$

Lemma 8.9. *u is the projection of x onto C iff $\langle x - u, y - u \rangle \leq 0 \, \forall y \in C$*

We have $f_i(x) = \|x - P_{C_i}(x)\|^2$ so $\partial f_i(x) = \frac{x - P_{C_i}(x)}{|x - P_{C_i}(x)|}$ if $x \neq 0$.

$$x_{t+1} = x_t - f(x_t) \cdot \frac{x_t - P_{C_i}(x_t)}{\|x_t - P_{C_i}(x_t)\|} = P_{C_i}(x_t).$$

$\square$

9   Feb 12

Definition 9.1. *For $\min_{x \in C} f(x)$ we define the projected subgradient method to be*

$$\begin{cases} y_{t+1} = x_t - \eta_t g_t \\ x_{t+1} = P_C(y_t) \end{cases}.$$

Theorem 9.2. $P_C(x) = \arg\min_{y \in C} ||x - y||^2$

Proof. We have $z = P_c(x)$ iff $\forall y \in C$, $\langle x - z, y - z \rangle \le 0$. Then for

$$\underbrace{-\nabla f(z)}_{x-z} \in \underbrace{N_c(z)}_{\{g: g^T(y-z) \le 0 \ \forall y \in C\}}.$$

so we have $(x - z)^T(y - z) \le 0$ $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Lemma 9.3. *Projections are contractions:*

$$|| \underbrace{P_C(x_1)}_{z_1} - \underbrace{P_C(x_2)}_{z_2} ||_2 \le ||x_1 - x_2||_2.$$

Proof. We have

1. $\langle x_1 - z_1, z_2 - z_1 \rangle \le 0 \forall z_2 \in C$

2. $\langle x_2 - z_2, z_1 - z_2 \rangle \le 0 \forall z_1 \in C$

3. Adding these two inequalities we have

$$\langle x_1 - z_1, z_2 - z_1 \rangle + \langle x_2 - z_2, z_1 - z_2 \rangle = \langle x_1 - x_2 + z_2 - z_1, z_2 - z_1 \rangle \le 0.$$

So

$$\langle x_1 - x_2, z_2 - z_1 \rangle + ||z_2 - z_1||_2^2 \le 0 \iff ||z_2 - z_1||_2^2 \le \langle x_1 - x_2, z_2 - z_1 \rangle \le ||x_1 - x_2||||z_2 - z_1||.$$

So we have our result

$$||z_1 - z_2|| \le ||x_1 - x_2||.$$

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Theorem 9.4. *Rates for PGD are identical to GD and similarly PSGM is identical to SGM.*

Proof.

$$||x_{t+1} - x^*||^2 = ||P_C(x_t - \eta_t g_t) - x^*||^2 = ||P_C(x_t - \eta_t g_t) - P_C(x^*)|| \le ||x_t - \eta_t g_t - x^*||^2.$$

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Example 9.5. *Consider*

$$\min_{x \in \mathbb{R}^d} \frac{1}{2} ||Ax - b||_2^2 \text{ where } n \ge d \text{ and } A \text{ full rank}.$$

When we solve this by gradient descent then the rate would be

$$\left(1 - \frac{\lambda_{min}}{\lambda_{max}}\right)^k.$$

Proof. We have $\nabla f(x) = A^T(Ax - b)$ and $\nabla^2 f(x) = A^T A$. So f is quadratic with Hessian $A^T A$. The eigenvalues of $A^T A$ satisfy $\lambda_{\min} = \lambda_{\min}(A^T A)$ and $\lambda_{\max} = \lambda_{\max}(A^T A)$. Hence f is α-strongly convex with $\alpha = \lambda_{\min}$ and β-smooth with $\beta = \lambda_{\max}$. Taking $\gamma = \beta/\alpha = \lambda_{\max}/\lambda_{\min}$ and $\eta = 1/\beta$, the strongly convex convergence theorem gives

$$||x_k - x^*||^2 \le \left(1 - \frac{1}{\gamma}\right)^k ||x_0 - x^*||^2 = \left(1 - \frac{\lambda_{\min}}{\lambda_{\max}}\right)^k ||x_0 - x^*||^2.$$

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Example 9.6. *Consider a similar problem*

$$\min_{x \geq 0} \frac{1}{2} \|Ax - b\|_2^2 \text{ where } n \geq d \text{ and } A \text{ full rank }.$$

We can let $y_{t+1} \leftarrow x_t - \eta_t(Ax_t - b)$ *and* $x_{t+1} \leftarrow \max(0, y_{t+1})$.

Example 9.7. *Example of Fast Projection:*

$$C = \left\{ Ay + b | y \in \mathbb{R}^\ell \right\} \subseteq \mathbb{R}^n \text{ and } P_C(x) = b + A(A^T A)^{-1} A^T (x - b).$$

Proof.

$$y^* = \arg\min_{y \in \mathbb{R}^\ell} \frac{1}{2} \|x - (Ay+b)\|_2^2 \implies A^T(Ay^*+b) - x = 0 \implies A^T Ay^* = A^T(x-b) \implies y^* = (A^T A)^{-1} A^T(x-b).$$

So $P_C(x) = Ay^* + b$ $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$