AIE6003: Optimization for Machine Learning

Xiaopeng Li
CUHK(SZ)
SAI

February 4, 2026

Outline

Subgradient Methods

Mirror Descent

Nonsmoothness in ML

Robust PCA:

$$\min_{X,Y} \quad \|M - XY^\top\|_1$$

Lasso Regression:

$$\min_{w\in\mathbb{R}^d} \quad \tfrac{1}{2n}\|Xw - y\|_2^2 + \lambda\|w\|_1$$

SVM:

$$\min_{w\in\mathbb{R}^d} \quad \tfrac{\lambda}{2}\|w\|_2^2 + \tfrac{1}{n}\sum_{i=1}^{n}\max(0, 1 - y_i\langle w, x_i\rangle).$$

ReLU Neural Network: $\sigma(z) := \max(0, z)$,

$$\min_{W_1,\ldots,W_n} \quad \frac{1}{n}\sum_{i=1}^{n}\left\|W_n\sigma\big(W_{n-1}\sigma(\cdots\sigma(W_1x_i)\cdots)\big) - y_i\right\|_F^2$$

Constraints:

$$\min_{w\in\mathbb{R}^d} \quad f(w) + \iota_{\mathcal{C}}(w), \quad \iota_{\mathcal{C}}(w) = \begin{cases} 0, & w \in \mathcal{C}, \\ +\infty, & w \notin \mathcal{C}. \end{cases}$$

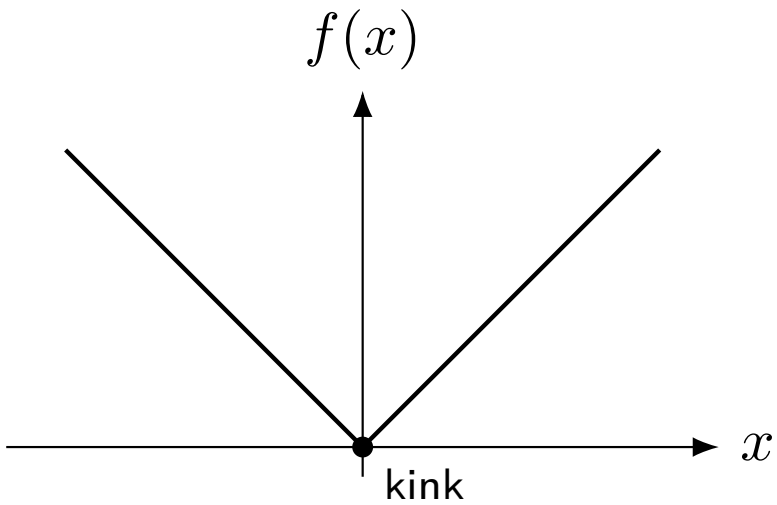
Nonsmooth Minimizer

Toy example:

$$f(x) = |x| = \begin{cases} x & x \geq 0 \\ -x & x < 0 \end{cases}$$

Not differentiable at $x = 0$, but

- f is convex and easy to evaluate.
- The optimum is at the kink $x^\star = 0$.



Many ML optima live at nonsmooth points; requiring a gradient everywhere is too restrictive.

Subgradients and Subdifferential

Extended Real-Valued: Consider convex $f : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$.

- Proper: $f(x) < +\infty$ for some x, i.e., $\text{dom}(f) \neq \emptyset$.

Subgradient: A vector $g \in \mathbb{R}^d$ is a subgradient of f at x if

$$f(y) \geq f(x) + \langle g, y - x \rangle \quad \forall y \in \mathbb{R}^d.$$

Subdifferential: the set of all subgradients: $\partial f(x)$.

- If f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$.
- If $\partial f(x)$ is a singleton, then f is differentiable at x and $\partial f(x) = \{\nabla f(x)\}$.

Properties:

- $\partial f(x)$ is closed and convex if f is convex.
- $\partial f(x) \neq \emptyset$ if f is proper convex and $x \in \text{int}(\text{dom}(f))$.
- $0 \in \partial f(x^*) \iff x^*$ is a global minimizer of f if f is convex.

Proof of Nonemptyness

Separation

Let $\mathcal{X}$ be closed and convex, and let $x \notin \mathcal{X}$. Then, there exists $s \neq 0$ such that $\langle s, x\rangle > \sup_{y\in\mathcal{X}}\langle s, y\rangle$.

Proof: set $s = x - \Pi_{\mathcal{X}}(x) \neq 0$.

- Variational characterization of projection gives $\forall\, y \in \mathcal{X}$,

$$\begin{aligned}\langle x - \Pi_{\mathcal{X}}(x), y - \Pi_{\mathcal{X}}(x)\rangle \leq 0 &\iff \langle s, y - (x - s)\rangle \leq 0 \\ &\iff \langle s, x\rangle \geq \langle s, y\rangle + \|s\|^2.\end{aligned}$$

- Taking $\sup$ on both sides yield:

$$\langle s, x\rangle \geq \sup_{y\in\mathcal{X}}\langle s, y\rangle + \|s\|^2 > \sup_{y\in\mathcal{X}}\langle s, y\rangle.$$

Proof of Nonemptyness

Local Lipschitzness

Let $f : \mathcal{X} \mapsto \mathbb{R}$ be convex and $y \in \text{int}(\mathcal{X})$, then there are L, ρ such that

$$|f(u) - f(v)| \leq L\|u - v\|, \quad \forall\, u, v \in B_\rho(y).$$

Localization: $E := \{(z, t) : z \in B_{r/2}(y), t \geq f(z)\}$ is closed and convex.

Separation: take $x_\delta = (y, f(y) - \delta) \notin E$, applying separation:

$$a^\top (z - y) + b(t - (f(y) - \delta)) \leq 0.$$

where we need to show

- $b < 0$
- $f(z) \geq f(y) - \delta + \langle g_\delta, z - y \rangle$, for all $z \in B_{r/2}(y)$.

Limit: g_δ is bounded, so taking $\delta \to 0$ yields local subgradient inequality.

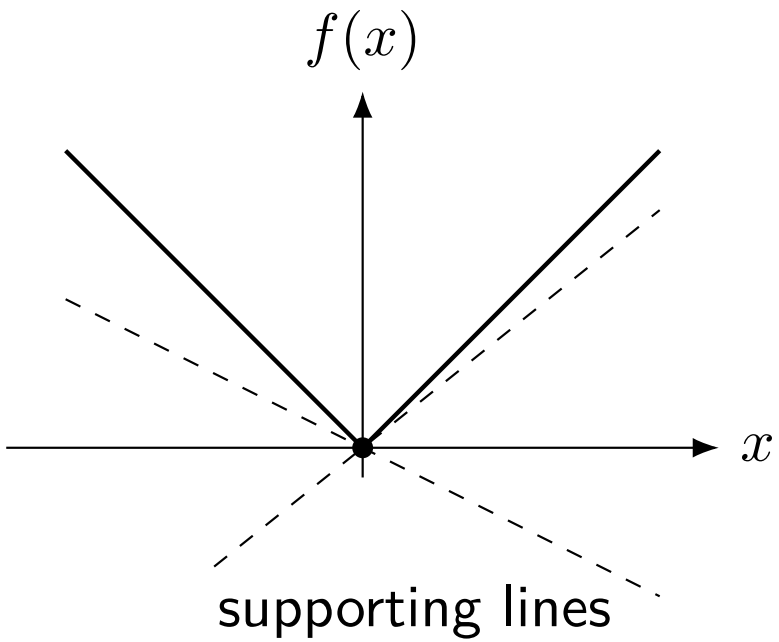
Extension: use convexity to extend local to global subgradient inequality.

Geometric intuition

Toy example: $f(x) = |x|$. For $x \neq 0$:

$$\partial f(x) = \begin{cases} \{\text{sign}(x)\}, & \text{if } x \neq 0 \\ [-1, 1] & \text{if } x = 0 \end{cases}$$

- Subgradients give global linear lower bounds.
- Analogy of $\nabla f(x^*) = 0$ without smoothness for minimizer x^*.
- Analogy of GD algorithm without smoothness.



Subdifferential Calculus

Assumption: $f : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ is proper convex.

Positive Scaling: $\partial(af) = a(\partial f)$ if $a > 0$.

Additivity: $\partial(f + g) = \partial f + \partial g$ if $\mathsf{int}(\mathsf{dom}(f)) \cap \mathsf{dom}(g) \neq \emptyset$.

Composition: for $g(x) := f(Ax)$, we have $\partial g(x) = A^\top \partial f(Ax)$ if $\mathsf{int}(\mathsf{dom}(f)) \cap \mathsf{range}(A) \neq \emptyset$.

Pointwise Maximum: for $f(x) := \max_{i=1}^m f_i(x)$, we have

$$\partial f(x) = \mathsf{conv}\left(\bigcup_{i: f_i(x) = f(x)} \partial f_i(x) \right),$$

where f_i's are additionally l.s.c., i.e., $\liminf_{y \to x} f_i(y) \geq f_i(x)$.

Convex Hull

Given $x_1, \ldots, x_m \in \mathbb{R}^n$, a convex combination is

$$x = \sum_{i=1}^{m} \lambda_i x_i, \quad \lambda_i \geq 0, \quad \sum_{i=1}^{m} \lambda_i = 1.$$

Convex hull: For a set $S \subseteq \mathbb{R}^n$, its convex hull is

$$\mathrm{conv}(S) := \left\{ \sum_{i=1}^{m} \lambda_i x_i : m \in \mathbb{N},\ x_i \in S,\ \lambda_i \geq 0,\ \sum_{i=1}^{m} \lambda_i = 1 \right\}.$$

Geometric intuition: $\mathrm{conv}(S)$ is the "smallest convex set containing S".

Property: If C is convex and $S \subseteq C$, then $\mathrm{conv}(S) \subseteq C$.

More Examples

Indicator Function: let $\mathcal{C}$ be a closed convex set, and define $\iota_{\mathcal{C}}$ as

$$\iota_{\mathcal{C}}(x) = \begin{cases} 0, & x \in \mathcal{C}, \\ +\infty, & x \notin \mathcal{C}. \end{cases}$$

Then $\partial \iota_{\mathcal{C}}(x) = N_{\mathcal{C}}(x)$, where the normal cone of $\mathcal{C}$ is

$$N_{\mathcal{C}}(x) := \{g : \langle g, y - x \rangle \leq 0, \ \forall y \in \mathcal{C}\}.$$

Support Function: for a bounded convex set $S \neq \emptyset$, define

$$\sigma_S(x) := \sup_{y \in S} \langle y, x \rangle,$$

where σ_S is the support function of S.

- σ_S is convex.
- $\partial \sigma_S(x) = \{y : \sigma_S(x) = \langle y, x \rangle\}$.

More Examples

ℓ_p-Norm: $\|x\|_p = \sigma_S(x)$ with $S = \{x : \|x\|_q \le 1\}$ where p and q are conjugate $1/p + 1/q = 1$.

- $\partial(\|x\|_p) = \{g : \|g\|_q \le 1, \langle g, x\rangle = \|x\|_p\}$
- $p \in (1, \infty)$:

$$\partial(\|x\|_p) = \begin{cases} \frac{\mathsf{sign}(x) \odot |x|^{p-1}}{\|x\|_p^{p-1}} & \text{if } x \neq 0 \\ \{g : \|g\|_q \le 1\} & \text{if } x = 0 \end{cases}$$

- $p = 1$:

$$g \in \partial(\|x\|_1) \iff g_i = \begin{cases} \mathsf{sign}(x_i) & \text{if } x_i \neq 0 \\ \in [-1, 1] & \text{if } x_i = 0 \end{cases}.$$

- $p = \infty$: let $I(x) := \{i : |x_i| = \|x\|_\infty\}$. Then

$$\partial(\|x\|_\infty) = \begin{cases} \mathsf{conv}\{\mathsf{sign}(x_i) e_i : i \in I(x)\} & \text{if } x \neq 0 \\ \{g : \|g\|_q \le 1\} & \text{if } x = 0 \end{cases}$$

Lasso Problem: A Special Case

Formulation: given $\lambda > 0$:

$$\min_{\beta} \quad f(\beta) := \tfrac{1}{2}\|y - \beta\|_2^2 + \lambda\|\beta\|_1$$

Optimality Condition: $0 \in \partial f(\beta^*)$

$$0 \in (\beta^\star - y) + \lambda\partial\|\beta^\star\|_1 \iff y - \beta^\star \in \lambda\partial\|\beta^\star\|_1.$$

Subgradient: $v \in \partial\|\beta\|_1$ is given by

$$v_i = \begin{cases} \text{sign}(\beta_i), & \beta_i \neq 0, \\ \in [-1, 1], & \beta_i = 0. \end{cases}$$

Optimal Solution: soft-thresholding operator $S_\lambda(\cdot)$

$$\beta_i^* = [S_\lambda(y)]_i := \begin{cases} y_i - \lambda, & y_i > \lambda, \\ 0, & |y_i| \leq \lambda, \\ y_i + \lambda, & y_i < -\lambda \end{cases}$$

Subgradient Methods

Goal: minimizing a convex function f with $\mathrm{dom}(f) = \mathbb{R}^n$.

Update: $x_{k+1} \in x_k - \alpha_k g_k$, where $g_k \in \partial f(x_k)$.

Subgradient Methods for L-Lipschitz Convex Functions

Suppose that f is convex with $\mathrm{dom}(f) = \mathbb{R}^n$ and f is L-Lipschitz. Let x^* be a minimizer of f and $R := \|x_0 - x^*\|$, then

$$e_k := \min_{i=1,\ldots,k} f(x_i) - f(x^*) \leq \frac{R^2 + L^2 \sum_{i=1}^k \alpha_i^2}{2 \sum_{i=1}^k \alpha_i}.$$

Step Sizes:

- constant step size $\alpha_k = \alpha$: $e_k = \frac{R^2}{2\alpha k} + \frac{\alpha L^2}{2} \to \frac{\alpha L^2}{2}$ as $k \to \infty$.
- nonsummable diminishing step size $\alpha_k = \frac{c}{\sqrt{k}}$: $e_k = O(\frac{\log k}{\sqrt{k}})$.
- Polyak step size $\alpha_k = \frac{f(x_k) - f(x^*)}{\|g_k\|^2}$: $e_k = O(\frac{1}{\sqrt{k}})$. Optimal rate!

Subgradient Methods

Update: $x_{k+1} \in x_k - \alpha_k g_k$, where $g_k \in \partial f(x_k)$.

The proof is similar to GD under for convex L-Lipschitz.

$$\|x_k - x^*\|^2 \leq \|x_{k-1} - x^*\|^2 - 2\alpha_{k-1}(f(x_{k-1}) - f(x^*)) + \alpha_{k-1}^2 L^2.$$

Telescoping:

$$2\sum_{i=1}^{k} \alpha_i (f(x_i) - f(x^*)) \leq R^2 + \sum_{i=1}^{k} \alpha_i^2 \|g_i\|^2.$$

Polyak Step Size:

$$\sum_{i=1}^{k} \frac{(f(x_i) - f(x^*))^2}{\|g_i\|^2} \leq R^2 \implies \sum_{i=1}^{k} (f(x_i) - f(x^*))^2 \leq L^2 R^2.$$

Outline

A Simplex Constrained Problem

Formulation: $\min_{x \in \Delta_d} f(x)$, where Δ_d is a d-simplex:

$$\Delta_d := \left\{ x \in \mathbb{R}^d,\ x_i \geq 0,\ \sum_{i=1}^{d} x_i = 1 \right\}.$$

Assumption: $f : \mathbb{R}^d \mapsto \mathbb{R}$ is convex differentiable and $\|\nabla f(x)\|_\infty \leq L$. There exists a minimizer x^*.

Naive Algorithm: Projected Gradient Descent (PGD)

$$x_{k+1} = \Pi_{\Delta_d}(x_k - \alpha_k \nabla f(x_k)).$$

Guarantee: fixed T, for $\bar{x}_T := \frac{1}{T}\sum_{i=1}^{T} x_i$ and $\alpha_k = \frac{\|x_1 - x^*\|_2}{L\sqrt{dT}}$,

$$f(\bar{x}_T) - f(x^*) \leq \frac{\|x_1 - x^*\|_2 L\sqrt{d}}{\sqrt{T}},$$

Fact: $\|x_1 - x^*\|_2 \leq \sqrt{2}$.

A Simple Fact

Fact: $\|x_1 - x^*\|_2 \leq \sqrt{2}$.

Proof:

- Δ_d has $d+1$ vertices $x_0, \ldots, x_d$.
- each point x in Δ_d can be written as $x = \sum_{i=0}^{d} \lambda_i x_i$ for some convex combination given by λ_i.
- We have

$$\|x - y\| = \left\| \sum_{i=0}^{d} \lambda_i x - \sum_{i=0}^{d} \lambda_i x_i \right\| = \left\| \sum_{i=0}^{d} \lambda_i (x - x_i) \right\|$$

$$\leq \sum_{i=0}^{d} \lambda_i \cdot \max_{i=0,\ldots,d} \|x - x_i\| = \max_{i=0,\ldots,d} \|x - x_i\|.$$

- Similarly, we have

$$\|x - x_i\| \leq \max_{i,j=0,\ldots,d} \|x_j - x_i\| = \sqrt{2}.$$

Proof

Update: $x_{k+1} = \Pi_{\Delta_d}(x_k - \alpha_k \nabla f(x_k))$.

Projection: using nonexpensiveness of projection,

$$\begin{aligned}\|x_{k+1} - x^*\|_2^2 &= \|\Pi_{\Delta_d}(x_k - \alpha_k \nabla f(x_k)) - x^*\|_2^2 \\ &\leq \|x_k - \alpha_k \nabla f(x_k) - x^*\|_2^2 \\ &= \|x_k - x^*\|_2^2 + \alpha_k^2 \|\nabla f(x_k)\|_2^2 - 2\alpha_k \langle \nabla f(x_k), x_k - x^* \rangle.\end{aligned}$$

Convexity and Telescoping:

$$2\sum_{k=1}^{T} \alpha_k (f(x_k) - f(x^*)) \leq \|x_1 - x^*\|_2^2 + \sum_{k=1}^{T} \alpha_k^2 \|\nabla f(x_k)\|_2^2.$$

Bound the gradient: $\|\nabla f(x)\|_2 \leq \sqrt{d}\|\nabla f(x)\|_\infty \leq \sqrt{d}\, L$.

$$f(\bar{x}_T) - f(x^*) \leq \frac{\|x_1 - x^*\|_2^2}{2\alpha T} + \frac{\alpha\, d L^2}{2}.$$

Step size: use $\alpha = \frac{\|x_1 - x^*\|_2}{L\sqrt{dT}}$ to get the desired result.

Exponentiated Gradient Descent (EGD)

Goal: minimizing a convex differentiable function on a d-simplex.

EGD: for $k \geq 1$,

$$x_{k+1,i} = \frac{x_{k,i} \exp\big(- \alpha_k \left[\nabla f(x_k)\right]_i\big)}{\sum_{j=1}^d x_{k,j} \exp\big(- \alpha_k \left[\nabla f(x_k)\right]_j\big)}, \qquad i = 1, \ldots, d.$$

Initialization: $x_1 = (1/d, \ldots, 1/d)$.

Convergence of EGD

Suppose f is convex differentiable and $\|\nabla f(x)\|_\infty \leq L$. Fixed T, for $\bar{x}_T := \frac{1}{T}\sum_{k=1}^T x_k$ and $\alpha_k \equiv \alpha = \frac{\sqrt{2\log d}}{L\sqrt{T}}$, we have

$$f(\bar{x}_T) - f(x^*) \leq L\sqrt{\frac{2\log d}{T}}.$$

Proof

KL divergence: $D_{\mathrm{KL}}(x^* \| x) := \sum_{i=1}^d x_i^* \log \frac{x_i^*}{x_i}$.

One-step progress:

$$\alpha \langle g_k, x_k - x^* \rangle \leq D_{\mathrm{KL}}(x^* \| x_k) - D_{\mathrm{KL}}(x^* \| x_{k+1}) + \frac{\alpha^2}{2} \|g_k\|_\infty^2.$$

Convexity and Telescoping:

$$\alpha \sum_{k=1}^T \left(f(x_k) - f(x^*) \right) \leq D_{\mathrm{KL}}(x^* \| x_1) + \frac{\alpha^2 L^2 T}{2}.$$

Average iterate: $f(\bar{x}_T) \leq \frac{1}{T} \sum_{k=1}^T f(x_k)$, so

$$f(\bar{x}_T) - f(x^*) \leq \frac{D_{\mathrm{KL}}(x^* \| x_1)}{\alpha T} + \frac{\alpha L^2}{2}.$$

KL Bound: $D_{\mathrm{KL}}(x^* \| x_1) \leq \log d$. Choose $\alpha = \frac{\sqrt{2 \log d}}{L\sqrt{T}}$ to obtain

$$f(\bar{x}_T) - f(x^*) \leq L \sqrt{\frac{2 \log d}{T}}.$$

Mirror Descent

Bregman Divergence: given a strictly convex $R : \mathcal{X} \mapsto \mathbb{R}$ on convex $\mathcal{X}$,

$$D_R(x, y) := R(x) - R(y) - \langle \nabla R(y), x - y \rangle.$$

Properties:

- $D_R(x, y) \geq 0$ for all x, y and $D_R(x, y) = 0 \iff x = y$.
- cosines law (3-point identity)

$$D_R(x, y) = D_R(x, z) + D_R(z, y) - \langle x - z, \nabla R(y) - \nabla R(z) \rangle.$$

- $D_R(x, y)$ is symmetric iff $D_R(x, y) = (x - y)^\top A(x - y)$, $A \succ 0$.
- $D_R(x, y)$ is convex in x but not necessarily convex in y.

Mirror Descent:

$$x_{k+1} = \arg\min_{x \in \mathcal{X}} \{D_R(x, x_k) + \alpha \langle g_k, x \rangle\},$$

where $\mathcal{X}$ is the constraint set and $g_k \in \partial f(x_k)$.

An Equivalent Form

Why mirror descent called "mirror" descent?

- x_k is mapped to the dual space (mirror) $\nabla R(x_k)$;
- gradient descent in the dual space: $y_{k+1} = \nabla R(x_k) - \alpha_k g_k$;
- y_{k+1} is mapped back to the original space $x'_{k+1} = (\nabla R)^{-1}(y_{k+1})$;
- project back to $\mathcal{X}$: $x_{k+1} = \arg\min_{z \in \mathcal{X}} D_R(z, x'_{k+1})$.

To verify equivalence, denote $w_{k+1} = (\nabla R)^{-1}(\nabla R(x_k) - \alpha_k g_k)$. Then

$$
\begin{aligned}
x_{k+1} &= \arg\min_{x \in \mathcal{X}} \; D_R(x, x_k) + \alpha \langle g_k, x \rangle, \\
&= \arg\min_{x \in \mathcal{X}} \; R(x) - R(x_k) - \langle \nabla R(x_k), x - x_k \rangle + \alpha \langle g_k, x \rangle \\
&= \arg\min_{x \in \mathcal{X}} \; R(x) - \langle \nabla R(w_{k+1}), x \rangle \\
&= \arg\min_{x \in \mathcal{X}} \; R(x) - R(w_{k+1}) - \langle \nabla R(w_{k+1}), x - w_{k+1} \rangle \\
&= \arg\min_{x \in \mathcal{X}} \; D_R(x, w_{k+1})
\end{aligned}
$$

Theory

Dual Space: $\mathcal{X}$ is the primal space, then the dual space is:

$$\mathcal{X}^* := \left\{ g : \|g\|_* := \sup_{\|x\| \leq 1} |\langle x, g \rangle| < \infty \right\}.$$

Example: $\mathcal{X} = \mathbb{R}^d$ with norm $\|\cdot\|$ then $\mathcal{X}^* = \mathbb{R}^d$ with norm $\|\cdot\|_*$.

Convergence of Mirror Descent for Convex Function

Let $f : \mathcal{X} \mapsto \mathbb{R}$ and $R : \mathcal{Y} \mapsto \mathbb{R}$ be convex with $\mathcal{X} \subseteq \mathcal{Y} \subseteq \mathbb{R}^d$. Assume ∇R is a bijection, $\|\nabla f(x)\|_* \leq G$ for all $x \in \mathcal{X}$, and R is σ-strongly convex w.r.t. $\|\cdot\|$. Then with $\alpha = \sqrt{\frac{2\sigma}{TG^2D}}$ and $D = D_R(x^*, x_0)$,

$$f(\bar{x}_T) - f(x^*) \leq G\sqrt{\frac{2D}{\sigma T}},$$

where $\bar{x}_T = \frac{1}{T}\sum_{i=1}^{T} x_i$.

Theory

One-step progress by Generalized Pythagorean Theorem:

$$\alpha\langle g_k, x_k - x^*\rangle \leq D_R(x^*, x_k) - D_R(x^*, x_{k+1}) + D_R(x_k, w_{k+1}) - D_R(x_{k+1}, w_{k+1}).$$

Smoothness in dual norm: if R is σ-strongly convex, then

$$D_R(x_k, w_{k+1}) - D_R(x_{k+1}, w_{k+1}) \leq \frac{\alpha^2}{2\sigma}\|g_k\|_*^2.$$

Convexity and Telescoping:

$$\alpha\sum_{k=0}^{T-1}\left(f(x_k) - f(x^*)\right) \leq D_R(x^*, x_0) + \frac{\alpha^2}{2\sigma}\sum_{k=0}^{T-1}\|g_k\|_*^2.$$

Average iterate: $f(\bar{x}_T) \leq \frac{1}{T}\sum_{k=0}^{T-1} f(x_k)$, so

$$f(\bar{x}_T) - f(x^*) \leq \frac{D_R(x^*, x_0)}{\alpha T} + \frac{\alpha}{2\sigma T}\sum_{k=0}^{T-1}\|g_k\|_*^2.$$

Generalized Pythagorean Theorem

Pythagorean Theorem in Euclidean Space:

$$\|z - x\|^2 \geq \|z - x^*\| + \|x^* - x\|^2, \quad z \in \mathcal{X}, \quad x^* = \Pi_{\mathcal{X}}(x).$$

Generalized Pythagorean Theorem for Bregman Divergence:

$$D_R(z, x) \geq D_R(z, x^*) + D_R(x^*, x), \quad z \in \mathcal{X}, \quad x^* = \Pi_{\mathcal{X},R}(x).$$

Proof:

- optimality condition for constrained convex optimization
- $\nabla_z D_R(z, x)\Big|_{z=x^*} = \nabla R(x^*) - \nabla R(x)$
- use cosine law (3-point identity)

Unified Framework

Choice of R: different R recover different first-order algorithms. We should adjust gradient update to fit problem geometry.

Subgradient Method: take $R(x) = \frac{1}{2}\|x\|_2^2$ so $D_R(x\|y) = \frac{1}{2}\|x-y\|_2^2$, the update becomes

$$\begin{aligned} x_{k+1} &= \arg\min_{x\in\mathcal{X}} \left\{ \langle \nabla f(x_k), x - x_k\rangle + \tfrac{1}{2\alpha}\|x - x_k\|^2 \right\} \\ &= \Pi_{\mathcal{X}}(x_k - \alpha g_k). \end{aligned}$$

Exponential Gradient Method: take $\mathcal{X} = \Delta_d$ and $R(x) = \sum_{i=1}^d x_i \log x_i$, so $D_R(x^*\|x) = D_{\mathrm{KL}}(x^*\|x)$. The update becomes multiplicative:

$$x_{k+1,i} = \frac{x_{k,i}\exp(-\alpha g_{k,i})}{\sum_{j=1}^d x_{k,j}\exp(-\alpha g_{k,j})}.$$

Wrap-up and Next Week

Week 4: Key takeaways

- Subgradient and subdifferential calculus.
- Subgradient method has $O(1/\sqrt{T})$-type guarantees.
- Geometry matters on the simplex: PGD gives a $\sqrt{d}$ dependence, while EGD improves this to $\sqrt{\log d}$.
- Bregman divergence and its property.
- Mirror descent and convergence guarantee.

Thank you for your attendance!

Questions?

Next week (Week 5): Proximal Operators & Proximal GD

References