Chapter 6 Subgradient Methods

Subgradient and Subdifferential

- Definition (Subgradient) Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be convex. A vector $g \in \mathbb{R}^n$ is a *subgradient* of f at a point $x \in \mathbf{dom}(f)$ if $f(y) \geq f(x) + g^\top (y - x), \forall y \in \mathbf{dom}(f)$.
  - The set of all subgradient at x is called the *subdifferential* of f at x denoted as ∂f.
- Lemma 6.2 (uniqueness when differentiable) If f is convex and differentiable at $x \in \mathbf{dom}(f)$, then $\partial f = \{\nabla f(x)\}$.
  - Proof
    - Let $y = x + \epsilon d$, if $g \in \partial f$, then $f(x + \epsilon d) \geq f(x) + \epsilon g^\top d$, then $\frac{f(x+\epsilon d) - f(x)}{\epsilon} \geq g^\top d, \forall d, \forall \epsilon$.
    - Letting $\epsilon \to 0$, then we have $\nabla f(x)^\top d \geq g^\top d, \forall d$. This only holds when $g = \nabla f(x)$.
    - If f convex, we can check the definition of subgradient holds.
  - Or can use Taylor expanssion and get $(\mathbf{g} - \nabla \mathbf{f}(\mathbf{x}))^\top (\mathbf{y} - \mathbf{x}) \leq \mathbf{r}(\mathbf{y} - \mathbf{x})$, then let $\mathbf{y} = \mathbf{x} + \epsilon(\mathbf{g} - \nabla \mathbf{f}(\mathbf{x}))$.
- Lemma 6.3 If we don't assume convexity, we can only say $\partial f(x) \subseteq \{\nabla f(x)\}$.
- Examples 6.4
  i. $f(x) = \frac{1}{2}x^2, \partial f(x) = x$
  ii. $f(x) = |x|, \partial f(x) = \begin{cases} \mathbf{sgn}(x), x \neq 0 \\ [-1, 1], x = 0 \end{cases}$
  iii. $f(x) = \begin{cases} -\sqrt{x}, x \geq 0 \\ +\infty, o.w. \end{cases}, \partial f(0) = \emptyset;$
  iv. $f(x) = \begin{cases} 1, x = 0 \\ 0, x > 0 \\ +\infty, o.w \end{cases}, \partial f(0) = \emptyset$
- Lemma 6.5 Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex, $\mathbf{dom}(f)$ open, $B \in \mathbb{R}_+$. Then $\forall x \in \mathbf{dom}(d), \forall g \in \partial f(x), \|g\|_{a*} \leq B \Leftrightarrow \forall x, y \in \mathbf{dom}(d), |f(x) - f(y)| \leq B\|x - y\|_a$.
  - Proof
    - (->)
      - Subgradient -> $f(y) - f(x) \leq g_y^\top (y - x)$ and $f(x) - f(y) \leq g_x^\top (y - x)$
      - This means $|f(y) - f(x)| \leq \mathbf{max}\{g_y^\top (y - x), g_x^\top (y - x)\} \leq \mathbf{max}\{\|g_y\|_{a*}, \|g_x\|_{a*}\}\|y - x\|_a$
      - If $\|g\|_{a*} \leq B$ we arrive at Lipschitz
    - (<-)
      - $f(y) \geq f(x) + g_x^\top (y - x)$, by Lipschitz, we have $B\|x - y\|_a \geq g_x^\top (y - x), \forall y$. Choosing $y - x = \mathbf{arg\,max}_{\|y-x\|=1} g_x^\top (y - x)$, we get maximum of $\|g_x\|_{a*} \leq B$. This is true forall x.

Topological Properties of Subgradients

- Lemma 6.6 Let $f(x)$ be a convex function and $x \in \mathbf{dom}(f)$. Then $\partial f(x)$ is convex and closed.
  - Proof $\partial f(x) = \cap_y \left\{g \in \mathbb{R}^n : f(y) \geq f(x) + g^\top (y - x)\right\}$, solution of g for each y is convex and closed linear separation.
- Definition 6.7 (separation of convex sets) Let S and T be two nonempty convex sets in $\mathbb{R}^n$. A hyperplane $H = \left\{x \in \mathbb{R}^n : a^\top x = b, a \neq 0\right\}$ is said to *separate* S and T if $S \cup T \not\subset H$ and $S \subset H^- = \left\{x \in \mathbb{R}^n : a^\top x \leq b\right\}$ and $T \subset H^+ = \left\{x \in \mathbb{R}^n : a^\top x \geq b\right\}$.
  - PS: It's OK that one of S or T is a point, or part of the hyperplane.
  - *Strictly separation* if $S \subset H^{--} = \left\{x \in \mathbb{R}^n : a^\top x < b\right\}$ and $T \subset H^{++} = \left\{x \in \mathbb{R}^n : a^\top x > b\right\}$.
- Theorem 6.8 (Hyperplane separation theorem, [Roc97, Thm 11.3], no proof) Let S and T be two nonempty convex sets. Then S and T can be separated if and only if their (relative) interiors do not intersect, $\mathbf{rint}(S) \cap \mathbf{rint}(T) = \emptyset$.
  - PS The *Relative Interior* of a general non=convex set is defined to be $\mathbf{relint}(C) := \{x \in C : \forall y \in C, \exists \lambda > 1, \text{ s.t. } \lambda x + (1 - \lambda) y \in C\}$
- Corollary 6.9 Let S be a nonempty convex set and $x_0 \in \partial S$. $\exists H = \left\{x : a^\top x = a^\top x_0\right\}, a \neq 0$ s.t. $S \subset \left\{x : a^\top x \leq a^\top x_0\right\}$, and $x_0 \in H$.
  - just let $T = \{x_0\}$.
- Theorem 6.10 (non emptiness of subgradient for interior) Let $f(x)$ be a convex function and $x \in \mathbf{rint}(\mathbf{dom}(f))$. Then $\partial f(x)$ is nonempty and bounded if $x \in \mathbf{rin}\, t(\mathbf{dom}(f))$.
  - Proof
    - *Non-emptiness*

        - W.l.o.g., assume $\mathbf{dom}(f)$ is full-dimensional and $x \in \mathbf{int}(\mathbf{dom}(f))$ (*interior*).
        - $\mathbf{epi}(f)$ is convex and $(x, f(x))$ belongs to its boundary, by Thm 6.8 $\exists \boldsymbol{\alpha} = (s, \beta) \neq 0$ s.t. $s^\top y + \beta t \geq s^\mathcal{T} x + \beta f(x), \forall (y,t) \in \mathbf{epi}(f)$
        - We claim $\beta > 0$.
            - If $\beta < 0$, we can let $t \to \infty$ where $(y,t) \in \mathbf{epi}(f)$.
            - if $\beta = 0$, this contradict to $s^\top y \geq s^\mathcal{T} x$, since x is the interior, we can always find opposite direction to reverse the inequality.
        - Setting $g = -\beta^{-1} s$, we have $f(y) \geq f(x) + g^\top (y - x), \forall y$
    - *Boundedness*
        - If $\partial f(x) \not< \infty$, $\exists \{g_k\}_{k=1}^\infty \in \partial f(x)$ s.t. $\lim_{k\to\infty} \|g_k\|_2 = \infty$.
        - Since x interior, exists ball $B(x, \delta) \subseteq \mathbf{dom}(f)$, then $y_k = x + \delta \frac{g_k}{\|g_k\|_2} \in \mathbf{dom}(f)$, by convexity, $f(y_k) \geq f(x) + g_k^\top (y_k - x) = f(x) + \delta \|g_k\|_2 \to \infty$, contradiction.
- Lemma A (non-emptyness -> convexity) If $\forall x \in \mathbf{dom}(f), \partial f(x) \neq \emptyset$ and $\mathbf{dom}(f)$ convex, then f convex.
    - Proof
        - $z = \lambda x + (1-\lambda) y \in \mathbf{dom}(f)$ for $\lambda \in [0,1]$.
        - Let $g \in \partial f(z)$, we have $f(x) \geq f(z) + g^\top (x - z), f(y) \geq f(z) + g^\top (y - z)$, add them with weight $(\lambda, 1-\lambda)$ we get convexity.
- Remark 6.11 (Exercise 42) The subdifferential of a convex function $f(x)$ at $x \in \mathbf{dom}(f)$ is a monotone operator, i.e., $(u - v)^\top (x - y) \geq 0, \forall x, y \in \mathbf{dom}(f), u \in \partial f(x), v \in \partial f(y)$. *Proof similar to gradient*.

Subdifferential and directional derivative

- Def directional derivative $f'(x; d) = \lim_{\delta \to 0^+} \frac{f(x+\delta d) - f(x)}{\delta}$.
- Def differentiable f *differentiable* if $\Delta f = A \Delta x + o(\Delta x)$.
    - ∂f can exists when $\mathbf{d} f$ not.
- Lemma 6.12 (Exercise 43) Let f convex, the ratio $\phi(\delta) := \frac{f(x+\delta d) - f(x)}{\delta}$ is non-decreasing of $\delta > 0$.
    - Proof Let $\delta_1 \geq \delta_2$, then let $\lambda := \delta_2 / \delta_1$, by convexity $\phi(\delta_2) = \frac{f(x + \lambda \delta_1 d) - f(x)}{\lambda \delta_1} \leq \frac{\lambda f(x + \delta_1 d) + (1-\lambda) f(x) - f(x)}{\lambda \delta_1} = \phi(\delta_1)$.
- Theorem 6.13 Let f be convex and $x \in \mathbf{int}(\mathbf{dom}(f))$, then $f'(x; d) = \mathbf{max}_{g \in \partial f(x)} g^\top d$.
    - Proof
        - By def of subgrad, we have $f'(x; d) \geq \mathbf{max}_{g \in \partial f(x)} g^\top d$.
        - Now we want to also prove the opposite.
            - Define $C_1 = \{(y,t) : f(y) < t\}$, (not $\mathbf{epi}(f)$), and $C_2(d) = \{(y,t) : y = x + \alpha d, t = f(x) + \alpha f'(x; d), \alpha \geq 0\}$.
                - so both C_1, C_2 convex, non-empty.
            - Since $\phi(\delta)$ non-decreasing, $(t >) f(x + \alpha d) \geq f(x) + \alpha f'(x; d), \forall \alpha \geq 0$, so $C_1 \cap C_2 = \emptyset$
            - By hyperplane separation theorem $\exists (g_0, \beta) \neq 0$ s.t. $g_0^\top (x + \alpha d) + \beta (f(x) + \alpha f'(x; d)) \leq g_0^\top y + \beta t, \forall \alpha \geq 0, \forall t > f(y)$
                - Similar to proof of Thm 6.10, we claim $\beta > 0$. Let $\tilde{g} = \beta^{-1} g_0$
                - we get $\tilde{g}^\top (x + \alpha d) + f(x) + \alpha f'(x; d) \leq \tilde{g}^\top y + f(y), \forall \alpha \geq 0$.
                - Setting $\alpha = 0$, we get $\tilde{g}^\top x + f(x) \leq \tilde{g}^\top y + f(y) \Leftrightarrow -\tilde{g} \in \partial f(x)$
                - setting $\alpha = 1, y = x$, we get $\tilde{g}^\top d + f'(x; d) \leq 0 \Leftrightarrow f'(x; d) \leq -\tilde{g}^\top d \leq \mathbf{max}_{g \in \partial f(x)} g^\top d$
        - This mean $f'(x; d) = \mathbf{max}_{g \in \partial f(x)} g^\top d$.
- Theorem B [Roc97, Theorem 25.5] (differentiable almost everywhere) A (proper) convex function f is differentiable almost everywhere on (the interior of) $\mathbf{dom}(f)$.
- Lemma C (optimal condition) If $0 \in \partial f(x)$, then x is a global minimum. Proof $f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^\top (\mathbf{y} - \mathbf{x}) = f(\mathbf{x})$.

Calculus of Subgradient

- Conic combination $h(x) = \lambda f(x) + \mu g(x)$, then $\partial h(x) = \lambda \partial f(x) + \mu \partial g(x), \forall x \in \mathbf{int}(\mathbf{dom}(h))$.
- affine composition $h(x) = f(Ax + b)$, then $\partial h(x) = A^T \partial f(Ax + b)$.
- supremum $h(x) = \mathbf{sup}_{\alpha \in \mathcal{A}} f_\alpha(x)$, then $\partial h(x) \supseteq \mathbf{conv} \{\partial f_\alpha(x) | \alpha \in \alpha(x)\}$
- superposition $h(x) = F(f_1(x), \ldots, f_m(x))$ where $F(y_1, \ldots, y_m)$ non-decreasing and convex, then $\partial h(x) \supseteq \left\{\sum_{i=1}^m d_i \partial f_i(x) : (d_1, \ldots, d_m) \in \partial F(y_1, \ldots, y_m)\right\}$.
- Example 6.14 Let $h(x) = \mathbf{max}_{y \in C} f(x, y)$ where $f(x,y)$ is convex in x for any y and C is closed. Then $\partial f(x, y_*(x)) \subset \partial h(x)$, where $y_*(x) = \mathbf{argmax}_{y \in C} f(x, y)$
    - Proof Let $g \in \partial f(x, y_*(x))$, then $h(z) \geq f(z, y_*(x)) \geq f(x, y_*(x)) + g^T (z - x) = h(z) + g^T (z - x)$

- Lemma 6.15 (Exercise 44) Consider the function $f(x) = \|x\|_a$. Then $\partial f(x) = \left\{g : g^{\mathcal{T}} x = \|x\|_a \ \wedge \ \|g\|_{a*} \leq 1\right\}$. (In particular $\partial f(0) = \{g : \|g\|_{a*} \leq 1\}$.
  - Proof
    - $g \in \partial f(x) \Leftrightarrow \forall \delta, \|x+\delta\|_a \geq \|x\|_a + g^\top \delta$
      - setting $\delta = -x$, we get $0 = \|0\|_a = \|x\|_a - g^\top x$.
      - By triangular ineq, $\|x\|_a + \|\delta\|_a \geq \|x+\delta\|_a \geq \|x\|_a + g^\top \delta \Rightarrow \|\delta\|_a \geq g^\top \delta, \forall \delta$.
        - subjecting to $\|\delta\|_a = 1$ we get $\|g\|_{a*} \leq 1$.

Subgradient Methods

- Remark 6.16 Negative subgradient may not be a descending direction.
  - *Example* $f(x_1, x_2) = |x_1| + 2|x_2|$ at $\mathbf{x} = (1,0), \partial f(\mathbf{x}) = \{(1, a) : a \in [-2, 2]\}$
    - $\mathbf{g} = (1,0), \mathbf{d} = -\mathbf{g}$ is descending direction, while $\mathbf{g} = (1,2), \mathbf{d} = -\mathbf{g}$ is not.
- Problem $\min f(x)$ s.t. $x \in X$. *Denote*:
  - $R := \sqrt{\max_{x,y \in X} \|x-y\|_2^2}$ the diameter of X.
  - $B := \sup_{x,y \in X} \frac{|f(x) - f(y)|}{\|x-y\|_2} < +\infty$ the Lipschitz constant under ℓ_2 norm.

Subgradient Descent

- Algorithm $x_{t+1} = \Pi_X \left(x_t - \gamma_t g\left(x_t\right)\right)$, where $g\left(x_t\right) \in \partial f\left(x_t\right)$.
- Lemma D(Descent Lemma) f convex, for any optimal $\mathbf{x}^* \in X^*$, $\left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|_2^2 \leq \left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2 - 2\gamma_t \left(f\left(\mathbf{x}_t\right) - f^*\right) + \gamma_t^2 \left\|\mathbf{g}_t\right\|_2^2$
  - Proof
    - $\left\|\mathbf{x}_{t+1} - \mathbf{x}^\star\right\|_2^2 = \left\|\Pi_X\left(\mathbf{x}_t - \gamma_t \mathbf{g}_t\right) - \mathbf{x}^\star\right\|_2^2$,
    - by Fact 4.1 (ii) RHS $\leq \left\|\mathbf{x}_t - \gamma_t \mathbf{g}_t - \mathbf{x}^\star\right\|_2^2 = \left\|\mathbf{x}_t - \mathbf{x}^\star\right\|_2^2 - 2\gamma_t \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}^\star \rangle + \gamma_t^2 \left\|\mathbf{g}_t\right\|_2^2$
    - By convexity $\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}^\star \rangle \leq f(\mathbf{x}_t) - f^\star$, then we arrive at conclusion.

$\mathcal{O}(1/\sqrt{t})$ convergence for B-Lipschitz convex functions

- Theorem 6.17 f convex, then subgradient descent gives
  $\max\{\min_{1 \leq t \leq T} f\left(\mathbf{x}_t\right), f\left(\hat{x}_T\right)\} - f^* \leq \frac{\left\|\mathbf{x}_1 - \mathbf{x}^*\right\|_2^2 + \sum_{t=1}^T \gamma_t^2 \left\|\mathbf{g}_t\right\|_2^2}{2\sum_{t=1}^T \gamma_t} \leq \frac{R^2 + \sum_{t=1}^T \gamma_t^2 B^2}{2\sum_{t=1}^T \gamma_t}$.
  - Proof
    - By descent lemma $\gamma_t \left(f\left(\mathbf{x}_t\right) - f^*\right) \leq \frac{1}{2}\left(\left\|x_t - x^*\right\|_2^2 - \left\|x_{t+1} - x^*\right\|_2^2 + \gamma_t^2 \left\|g\left(x_t\right)\right\|_2^2\right)$
    - sum over $t = [1:T]$ and divided by $\sum_t \gamma_t$ gives upper bound of the average$\mathbb{E} f_t$, after that is straightforward.
- Step size is crucial for convergence behavior, unlike GD
  - *Constant Stepsize* $\gamma_t \equiv \gamma$, $\epsilon_T \leq \frac{R^2}{2T} \cdot \frac{1}{\gamma} + \frac{B^2}{2}\gamma \overset{T \to \infty}{\longrightarrow} \frac{B^2}{2}\gamma$,
    - choosing $\gamma_* = \frac{R}{B\sqrt{T}}$ yields $\epsilon_T \leq \frac{RB}{\sqrt{T}}$.
  - *Scaled stepsize* $\gamma_t = \frac{\gamma}{\|\mathbf{g}_t\|_2}$, $\epsilon_T \overset{T \to \infty}{\longrightarrow} B\gamma/2$
  - *Non-summable but diminishing stepsize* $\sum_{t=1}^\infty \gamma_t = \infty, \lim_{t \to \infty} \gamma_t = 0$
    - Since γ_t^2 approach zero faster than γ_t, $\epsilon_T \overset{T \to \infty}{\longrightarrow} 0$.
    - *Example* $\gamma_t = O\left(t^{-q}\right)$ with $q \in (0, 1]$, e.g. $q = 1/2$, then $\epsilon_T \leq O\left(\frac{BR\ln(T)}{\sqrt{T}}\right)$
      - If we average among $[T/2 : T]$, then $\min_{\lfloor \frac{T}{2} \rfloor \leq t \leq T} f\left(x_t\right) - f^* \leq O\left(\frac{BR}{\sqrt{T}}\right)$
- *Non-summable but square-summable (Robbins-Monro) stepsize* $\sum_{t=1}^\infty \gamma_t = \infty, \sum_{t=1}^\infty \gamma_t^2 < \infty$, e.g. $\gamma_t = t^{-1}$, $\epsilon_T \overset{T \to \infty}{\longrightarrow} 0$.
- *Polyak stepsize* Assume $f^* = f\left(x^*\right)$ is known and $\gamma_t = \frac{f(x_t) - f^*}{\|g(x_t)\|_2^2}$
  - Using descent lemma, we get $\left\|x_{t+1} - x^*\right\|_2^2 \leq \left\|x_t - x^*\right\|_2^2 - \frac{\left(f(x_t) - f^*\right)^2}{\|g(x_t)\|_2^2}$
  - then $\left(f(x_t) - f(x^\star)\right)^2 \leq \|\mathbf{g}_t\|^2 (\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2) \leq B^2 (\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2)$
  - then $\sum_{t=1}^T \left(f\left(x_t\right) - f^*\right)^2 \leq R^2 \cdot M < \infty$, this means $f\left(\mathbf{x}_t\right) \to f^*$.
  - **Note** this is *last iterate* convergence, not *average* convergence!
  - can also show that $\exists \mathbf{x}^\star \in X^\star$ s.t. $\limsup_{t \to \infty} \left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2 \to 0$, converge to one of the optimal point.

- Since series $\|\mathbf{x}_t - \mathbf{x}^*\|_2$ is bounded and non-increasing, there exists a subsequence $\{\mathbf{x}_{t_k}\}$ with accumulation point $\hat{\mathbf{x}}$,
- Since we already have $f(\mathbf{x}_t) \to f^*$, then $\hat{\mathbf{x}} \in X^*$.
- So $\|\mathbf{x}_{t+1} - \hat{\mathbf{x}}\|_2^2 \le \|\mathbf{x}_{t_k} - \hat{\mathbf{x}}\|_2^2 - \sum_{j=t_k}^{t} \frac{(f(\mathbf{x}_j) - f^*)^2}{\|g(\mathbf{x}_j)\|_2^2}$,
- taking $\limsup$ we have $\limsup_{t\to\infty} \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 \le \|\mathbf{x}_{t_k} - \hat{\mathbf{x}}\|_2^2 - \sum_{j=t_k}^{\infty} \frac{(f(\mathbf{x}_j) - f^*)^2}{\|g(\mathbf{x}_j)\|_2^2}$
- then take $k \to \infty$, we get $\limsup_{t\to\infty} \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 \le \lim_{k\to\infty} \|\mathbf{x}_{t_k} - \hat{\mathbf{x}}\|_2^2 = 0$.

Similar result under Polyak's step size (Exercise Prob 5)

- For non-differentiable convex f, μ-strongly convex w.r.t. 2-norm is defined to be $f_\mu(x) := f(x) - \frac{\mu}{2}\|\mathbf{x}\|^2$ is convex. It can also be shown that this means $f(y) \ge f(x) + g^\top (y - x) + \frac{\mu}{2}\|x - y\|^2$.
- Lemma I If f convex and B-Lipschitz, optimization in whole space $x \in \mathbb{R}^d$ with Polyak step size gives $\min_{1\le t\le T} f(x_t) - f(x^*) \le \frac{B\|x_1 - x^*\|_2}{\sqrt{T}}$.
  - Proof
    - Denote $h_t := f(x_t) - f(x^*)$ and $d_t := \|x_t - x^*\|_2$, update gives $d_{t+1}^2 \le d_t^2 - \frac{h_t^2}{\|\mathbf{g}_t\|_2^2}$
    - equivalently $h_t^2 \le \|\mathbf{g}_t\|_2^2 \left(d_t^2 - d_{t+1}^2\right) \le B^2 \left(d_t^2 - d_{t+1}^2\right)$
    - sum over $t = [1:T]$ gives the result.
- Lemma J If f μ-strongly convex and B-Lipschitz, optimization in whole space $x \in \mathbb{R}^d$ with Polyak step size gives $\min_{1\le t\le T} f(x_t) - f(x^*) \le \frac{4B^2}{\mu T}$
  - Proof
    - Similarly, $d_{t+1}^2 - d_t^2 \le -\frac{h_t^2}{\|\mathbf{g}_t\|_2^2} \le -\frac{h_t^2}{B^2}$
    - By strong convexity w.r.t. optimial point (since optimization on whole space, $0 \in \partial f(\mathbf{x}^\star)$, $h_t = f(\mathbf{x}_t) - f(\mathbf{x}^\star) \le \mu d_t^2/2$,
      - this means $d_{t+1}^2 - d_t^2 \le -\frac{h_t^2}{B^2} \le -\frac{\mu^2}{4B^2} d_t^4$. Denote $a_t = \mu^2\|x_t - x^*\|_2^2/\left(4B^2\right)$, we have $a_{t+1} \le a_t(1 - a_t)$
    - By strong convexity w.r.t. x_0 and $x^\star$, we have $f^\star \ge f(x_0) + \mathbf{g}_0^\top(x^\star - x_0) + \mu\|\Delta x_0\|_2^2/2$
      - by sufficient descent of Polyak step size, $0 \ge -h_0 \ge \mu d_0^2/2 - \mathbf{g}_0^\top \Delta x_0$,
      - this means $\mu d_0^2/2 \ge \mathbf{g}_0^\top \Delta x_0 \le B d_0$, take square we get $a_0 \le 1$.
    - We claim $a_t \le 1/(t+1)$, by induction, we have $a_{t+1} \le \frac{1}{t+2}\frac{t(t+2)}{(t+1)^2}$ (by monotonicity of $x(1-x)$ in $[0, 1/2]$)
      - Since $b_t := t/t+1$ is increasing, $b_t/b_{t+1} = \frac{t(t+2)}{(t+1)^2} \le 1$, this means $a_{t+1} \le \frac{1}{t+2}$, induction succeeds.
      - This means $d_t^2 \le \frac{4B^2}{\mu^2}\frac{1}{t+1} \le \frac{4B^2}{\mu^2}\frac{1}{t}$
    - Again, by $d_{t+1}^2 - d_t^2 \le -\frac{h_t^2}{B^2}$, sum over $t \in [T/2, T]$, we get $\sum_{t=T/2}^{T} h_t^2 \le B^2(d_{T/2}^2 - d_T^2) \le B^2 d_{T/2}^2 \le \frac{8B^4}{\mu^2}\frac{1}{T}$
      - this means $\min_{t\in[T/2:T]} h_t^2 \le \frac{16B^4}{\mu^2}\frac{1}{T^2}$, taking square root and we finish our proof.

$\mathcal{O}(1/t)$ convergence for strong convex functions

- Lemma E(Descent Lemma under strong convexity) f is μ-strongly convex, then $\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 \le (1 - \mu\gamma_t)\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\gamma_t\left(f(\mathbf{x}_t) - f^*\right) + \gamma_t^2\|\mathbf{g}_t\|_2^2$
  - Proof Replace $\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}^\star\rangle \le f(\mathbf{x}_t) - f^\star$ by $\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}^\star\rangle \le f(\mathbf{x}_t) - f^\star - (\mu/2)\|\mathbf{x}_t - \mathbf{x}^\star\|_2^2$.
- There are two choice of step size $1/\mu t$ or $2/\mu(t+1)$, which give different convergence rate.
- Theorem 6.18 ($\mathcal{O}(\ln T/T)$) f is μ-strongly convex and B-Lipschitz, then subgradient descent of step size $\gamma_t = 1/\mu t$ gives $\max\{\min_{1\le t\le T} f(\mathbf{x}_t), f\left(\hat{x}_T\right)\} - f^* \le \frac{B^2(\ln(T)+1)}{2\mu T}$ where $\hat{x}_T := \frac{1}{T}\sum_{t=1}^{T} x_t$.
  - Proof
    - Plug in $\gamma = 1/\mu t$ we have $f(x_t) - f^* \le \left(\frac{\mu}{2}(t-1)\|x_t - x^*\|_2^2 - \frac{\mu}{2}t\|x_{t+1} - x^*\|_2^2 + \frac{1}{2\mu t}\|g(x_t)\|_2^2\right)$,
    - sum over $t = [1:T]$ we get $\sum_{t=1}^{T}[f(x_t) - f^*] \le \sum_{t=1}^{T}\frac{1}{2\mu t}\|g(x_t)\|_2^2 \le \frac{B^2}{2\mu}\sum_{t=1}^{T}\frac{1}{t} \le \frac{B^2}{2\mu}(\ln(T) + 1)$
- Theorem F ($\mathcal{O}(1/T)$) f is μ-strongly convex and B-Lipschitz, then subgradient descent of step size $\gamma_t = 2/\mu(t+1)$ gives $\max\{\min_{1\le t\le T} f(\mathbf{x}_t), f\left(\hat{x}_T\right)\} - f^* \le \frac{2B^2}{\mu\cdot(T+1)}$ where $\hat{x}_T := \sum_{t=1}^{T} t x_t / \sum_{t=1}^{T} t$.
  - Proof
    - By descent lemma $\left(f(\mathbf{x}_t) - f^*\right) \le \frac{1-\mu\gamma_t}{2\gamma_t}\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - \frac{1}{2\gamma_t}\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 + \frac{\gamma_t}{2}\|\mathbf{g}_t\|_2^2$
    - Plug in $\gamma_t = 2/\mu(t+1)$ we get $\mathbf{RHS} = \frac{\mu(t-1)}{4}\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - \frac{\mu(t+1)}{4}\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 + \frac{1}{\mu(t+1)}\|\mathbf{g}_t\|_2^2$
    - Times t on each side and we get $t\left(f(\mathbf{x}_t) - f^*\right) \le \frac{\mu t(t-1)}{4}\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - \frac{\mu(t+1)t}{4}\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 + \frac{t}{\mu(t+1)}\|\mathbf{g}_t\|_2^2$
    - Since $\frac{t}{\mu(t+1)}\|\mathbf{g}_t\|_2^2 \le \frac{1}{\mu}B^2$, sum over $t = [1:T]$ and get $\sum_{t=1}^{T} t\left(f(\mathbf{x}_t) - f^*\right) \le -\frac{\mu T(T+1)}{4}\|\mathbf{x}_{T+1} - \mathbf{x}^*\|_2^2 + \frac{T}{\mu}B^2$
    - else is straightforward.

Lower Bound Complexity

We can show that, the above $\mathcal{O}(1/\sqrt{t})$ and $\mathcal{O}(1/t)$ can not be improved. The worst case function is given by the piecewise-linear function similar to $f(x) = \max_{1\le i\le n} x_i$.

- Theorem 6.20 (Nemirovski & Yudin 1979). For any $1 \le t \le n, x_1 \in \mathbb{R}_n$,
  - (i) *B-Lipschit* there exists a B-Lipschitz continuous function f and a convex set X with diameter R, such that for any first-order method that generates $\mathbf{x}_t \in \mathbf{x}_1 + \mathbf{span}\left(\mathbf{g}_1, \ldots, \mathbf{g}_{t-1}\right)$, where $\mathbf{g}_i \in \partial f\left(\mathbf{x}_i\right)$, we have $\min_{1\le s\le t} f\left(\mathbf{x}_s\right) - f^* \ge \frac{B\cdot R}{4(1+\sqrt{t})}$.
  - (ii) *B-Lipschitz and μ-convex* there exists a μ-strongly convex, B-Lipschitz continuous function f and a convex set X with diameter R, for any first-order method as described above, we always have $\min_{1\le s\le t} f\left(x_s\right) - f^* \ge \frac{B^2}{8\mu t}$.
  - Proof
    - The function we construct is $f(x) = C\cdot \max_{1\le i\le t} x_i + \frac{\mu}{2}\|x\|_2^2$, and we restrict ourself within region $X = \{x \in \mathbb{R}^n, \|x\|_2 \le R\}$.
      - The subgradient is $\partial f(x) = \mu x + C\cdot \mathbf{conv}\left\{e_i : i \textbf{ such that } x_i = \max_{1\le j\le t} x_j\right\}$.
      - The optimal solution is $x_i^* = \begin{cases} -\frac{C}{\mu t} & 1\le i\le t \\ 0 & t < i \le n \end{cases}$ and $f^* = -\frac{C^2}{2\mu t}$, can be verified by $0 \in \partial f(x)$.
    - Since all subgradient method controls γ_T, but not how to choose subgradient, we can design a process that gives worst case. The update is design to have $g(x) = C\cdot e_i + \mu x$, only one pure direction of $x_i = \max_{1\le j\le t} x_j$, but not its combination.
      - To simplify, we always choose the smallest one.
    - We claim that if we start from $x_{s=1} = 0$, then $x_s \in \mathbf{span}\left(e_1, \ldots, e_{s-1}\right)$, always one dimension less than s.
      - When $s = 2$, all coordinates are equal and we choose e_1 to update, so $x_{s=2} \in \mathbf{span}\{e_1\}$.
      - By our update rule $g(x) = C\cdot e_i + \mu x$, the span can only increase at most one dimension at each update, so induction still holds.
    - This means for $x_s, 1 \le s \le t$, at least one coordinate is unchanged as 0. So $\max_{1\le i\le t} x_i$ -> $f(x_s) \ge 0$.
      - This gives $\min_{1\le s\le t} f\left(x_s\right) - f^* \ge \frac{C^2}{2\mu t}$
    - Choosing $C = \frac{B\sqrt{t}}{1+\sqrt{t}}, \mu = \frac{2B}{R(1+\sqrt{t})}$, we have $\|\partial f(x)\|_2 \le C + \mu\|x\|_2 \le C + \mu R =: M$, this is the M-Lipschitz case, we have $\min_{1\le s\le t} f\left(x_s\right) - f^* \ge \frac{C^2}{2\mu t} = \frac{B\cdot R}{2(1+\sqrt{t})}$.
    - Choosing $C = \frac{B}{2}, \mu = \frac{B}{R}$, we have $\|\partial f(x)\|_2 \le C + \mu R =: M$, this is M-Lispchitz and μ-strongly convex case, we have $\min_{1\le s\le t} f\left(x_s\right) - f^* \ge \frac{C^2}{2\mu t} = \frac{B^2}{8\mu t}$
  - *Comment* This theorem is so strange that number of updates is related to dimension.

Mirror Descent

- Key Idea Update can be viewed as $x_{t+1} = \mathop{\mathbf{argmin}}_{x\in X}\left\{\frac{1}{2}\|x - x_t\|_2^2 + \langle \gamma_t g\left(x_t\right), x\rangle\right\}$, how about we change $\|\cdot\|_2^2$ to something else.

Bregman Divergence

- Definition 6.22 Let $\omega(x) : X \to \mathbb{R}$ be a function that is *strictly* convex, continuously differentiable on a closed convex set X. The *Bregman divergence* is defined as $V_\omega(x, y) = \omega(x) - \omega(y) - \nabla\omega(y)^T(x - y), \forall x, y \in X$.
  - Asymmetric $V_\omega(x, y) \ne V_\omega(y, x)$, not a valid distance, triangle inequality may not hold.
  - $\omega(\cdot)$ is called *distance-generating function*.
  - If $\omega(\cdot)$ is also σ-strongly convex w.r.t. norm $\|\cdot\|_a$, then $V_\omega(x, y) \ge \sigma\|x - y\|_a^2/2$.
- Example
  - *Euclidean Distance* $\Omega = \mathbb{R}^d, \omega(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$, ω is 1 strongly convex w.r.t ℓ_2 norm,
    - $V_\omega(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$.
  - *Mahalanobis distance* $\Omega = \mathbb{R}^d, \omega(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q\mathbf{x}$ where $Q \succeq I$, ω is 1 strongly convex w.r.t ℓ_2 norm,
    - $V_\omega(\mathbf{x}, \mathbf{y}) = \frac{1}{2}(\mathbf{x} - \mathbf{y})^T Q(\mathbf{x} - \mathbf{y})$.
  - *Kullback-Leibler divergence* $\Omega = \Delta_d, \omega(\mathbf{x}) = \sum_{i=1}^d x_i \ln x_i$, $\nabla\omega(x) = \ln(x) - 1$,
    - $V_\omega(x, y) = \sum_i x_i \ln x_i - y_i \ln y_i - (\ln(y_i) - 1)(x_i - y_i) = \sum_i x_i \ln \frac{x_i}{y_i} = \mathbf{KL}(x\|y)$, since $\sum_i x_i = \sum_i y_i = 1$.
    - The proof for $V_\omega(x, y) \ge \|x - y\|_1^2/\ln 4$ is complicated, see this.
- Lemma 6.23 (Generalized Pythagorean Theorem, Exercise 46) If x^* is the Bregman projection of x_0 onto a convex set

$C \subset X : x^* \operatorname{argmin}_{x \in C} V_\omega(x, x_0)$. Then $\forall y \in C$, $V_\omega(y, x_0) \geq V_\omega(y, x^*) + V_\omega(x^*, x_0)$.

- Proof
  - $\nabla_x V_\omega(x, y) = \nabla\omega(x) - \nabla\omega(y)$, optimal condition means $\forall y \in C, (\nabla\omega(x^*) - \nabla\omega(x_0))^\top (y - x^*) \geq 0$.
  - equivalently, $-\nabla\omega(x_0)^\top (y - x_0 + x_0 - x^*) \geq -\nabla\omega(x^*)^\top (y - x^*)$
  - or $-\nabla\omega(x_0)^\top (y - x_0) \geq -\nabla\omega(x^*)^\top (y - x^*) - \nabla\omega(x_0)^\top (x_0 - x^*)$
  - then
    $\omega(x_0) - \omega_(y) - \nabla\omega(x_0)^\top (y - x_0) \geq \omega(x_0) - \omega_(x^*) + \omega(x^*) - \omega_(y) - \nabla\omega(x^*)^\top (y - x^*) + -\nabla\omega(x_0)^\top (x_0 - x^*)$,
    QED.

Mirror Descent Algorithm

- *Prox-mapping* $\mathbf{Prox}_x(\xi) = \underset{u \in X}{\operatorname{argmin}} \{V_\omega(u, x) + \langle \xi, u \rangle\}$, and suppose ω is 1-strongly convex on norm $\|\cdot\|_a$.
- Algorithm $x_{t+1} = \mathbf{Prox}_{x_t}(\gamma_t g(x_t))$, where $g(x_t) \in \partial f(x_t)$.
- Example 6.25 Under KL divergence, the prox-mapping becomes $\mathbf{Prox}_x \xi) = \left(\sum_{i=1}^n x_i e^{-\xi_i}\right)^{-1} \begin{bmatrix} x_1 e^{-\xi_1} \\ \dots \\ x_n e^{-\xi_n} \end{bmatrix}$
  - Proof
    - $V_\omega(u, x) + \langle \xi, u \rangle = \sum_i u_i \ln(u_i / x_i) + u_i \xi_i$, optimal condition under constraint is $0 = \ln(u_i / x_i) + 1 + \lambda + \xi_i$
    - So solution is $u_i = x_i e^{-\xi_i + \alpha}$ where $\sum_i u_i = 1 \Rightarrow e^{-\alpha} = \sum_i x_i e^{-\xi_i}$ QED.

$\mathcal{O}(1/\sqrt{t})$ convergence for convex functions

- Lemma 6.26 (Three point identity) $\forall x, y, z \in \mathbf{dom}(\omega)$, $V_\omega(x, z) = V_\omega(x, y) + V_\omega(y, z) - \langle \nabla\omega(z) - \nabla\omega(y), x - y \rangle$
  - Proof
    - $V_\omega(x, y) + V_\omega(y, z) = \omega(x) - \omega(y) + \omega(y) - \omega(z) - \langle \nabla\omega(y), x - y \rangle - \langle \nabla\omega(z), y - z \rangle$
    - RHS $= V_\omega(x, z) + \langle \nabla\omega(z), x - z \rangle - \langle \nabla\omega(y), x - y \rangle - \langle \nabla\omega(z), y - z \rangle$
    - RHS $= V_\omega(x, z) + \langle \nabla\omega(z) - \nabla\omega(y), x - y \rangle$
  - When $\omega = \|\cdot\|_2^2$, this becomes law of cosines.
- Lemma G (Descent Lemma) $\gamma_t (f(\mathbf{x}_t) - f^*) \leq V_\omega(\mathbf{x}^*, \mathbf{x}_t) - V_\omega(\mathbf{x}^*, \mathbf{x}_{t+1}) + \frac{\gamma_t^2}{2} \|\mathbf{g}_t\|_{a*}^2$
  - Proof
    - By 3-point identiy, $V_\omega(\mathbf{x}^\star, \mathbf{x}_t) = V_\omega(\mathbf{x}^\star, \mathbf{x}_{t+1}) + V_\omega(\mathbf{x}_{t+1}, \mathbf{x}_t) - \langle \nabla\omega(\mathbf{x}_t) - \nabla\omega(\mathbf{x}_{t+1}), \mathbf{x}^\star - \mathbf{x}_{t+1} \rangle$, equivalently
      - $V_\omega(\mathbf{x}^\star, \mathbf{x}_t) - V_\omega(\mathbf{x}^\star, \mathbf{x}_{t+1}) + \langle \nabla\omega(\mathbf{x}_t) - \nabla\omega(\mathbf{x}_{t+1}), \mathbf{x}^\star - \mathbf{x}_{t+1} \rangle = V_\omega(\mathbf{x}_{t+1}, \mathbf{x}_t)$
    - by def of algo, the optimial condition is $\forall x \in X$, especially $\mathbf{x}^\star$
      $\langle \nabla_{\mathbf{x}_{t+1}} V_\omega(\mathbf{x}_{t+1}, \mathbf{x}_t) + \gamma_t \mathbf{g}_t, \mathbf{x}^\star - \mathbf{x}_{t+1} \rangle = \langle \nabla\omega(\mathbf{x}_{t+1}) - \nabla\omega(\mathbf{x}_t) + \gamma_t \mathbf{g}_t, \mathbf{x}^\star - \mathbf{x}_{t+1} \rangle \geq 0$, this means
      - $V_\omega(\mathbf{x}^\star, \mathbf{x}_t) - V_\omega(\mathbf{x}^\star, \mathbf{x}_{t+1}) + \langle \gamma_t \mathbf{g}_t, \mathbf{x}^\star - \mathbf{x}_{t+1} \rangle \geq V_\omega(\mathbf{x}_{t+1}, \mathbf{x}_t)$
    - By convexity, $\langle \mathbf{g}_t, \mathbf{x}^\star - \mathbf{x}_t \rangle \leq f(\mathbf{x}^\star) - f(\mathbf{x}_t)$, so
      $V_\omega(\mathbf{x}^\star, \mathbf{x}_t) - V_\omega(\mathbf{x}^\star, \mathbf{x}_{t+1}) + \gamma_t (f(\mathbf{x}^\star) - f(\mathbf{x}_t)) \geq V_\omega(\mathbf{x}_{t+1}, \mathbf{x}_t) + \langle \gamma_t \mathbf{g}_t, \mathbf{x}_{t+1} - \mathbf{x}_t \rangle$
    - By *Young's inequality*, $\langle a, b \rangle \leq \|a\|_a \cdot \|b\|_{a*} \leq \|a\|_a^2 / 2 + \|b\|_{a*}^2 / 2$, therefore
      - $\langle \gamma_t \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \leq \gamma_t^2 \|\mathbf{g}_t\|_{a*}^2 / 2 + \|\mathbf{x}_t - \mathbf{x}_{t+1}\|_a^2 / 2$
    - By the assumption of ω is 1-strongly convex, $V_\omega(\mathbf{x}_{t+1}, \mathbf{x}_t) \geq \|\mathbf{x}_t - \mathbf{x}_{t+1}\|_a^2 / 2$
    - Combine the above two,
      $V_\omega(\mathbf{x}_{t+1}, \mathbf{x}_t) + \langle \gamma_t \mathbf{g}_t, \mathbf{x}_{t+1} - \mathbf{x}_t \rangle \geq \|\mathbf{x}_t - \mathbf{x}_{t+1}\|_a^2 / 2 - \gamma_t^2 \|\mathbf{g}_t\|_{a*}^2 / 2 - \|\mathbf{x}_t - \mathbf{x}_{t+1}\|_a^2 / 2 = -\gamma_t^2 \|\mathbf{g}_t\|_{a*}^2 / 2$
    - Combine above, we arrive at the descent lemma.
- Theorem 6.28 f convex, then $\max \left\{ \min_{1 \leq t \leq T} f(x_t), f\left(\hat{x}_T\right) \right\} - f^* \leq \frac{V_\omega(x^*, x_1) + \frac{1}{2} \sum_{t=1}^T \gamma_t^2 \|g(x_t)\|_*^2}{\sum_{t=1}^T \gamma_t}$, where $\hat{x}_T = \frac{\sum_{t=1}^T \gamma_t x_t}{\sum_{t=1}^T \gamma_t}$.
  - Proof Similar to Theorem 6.17, sum over all $t \in [1 : T]$ we get the result.
  - Convergence similar to subgradient case $\min_{1 \leq t \leq T} f(\mathbf{x}_t) - f^* = O\left(\frac{BR}{\sqrt{T}}\right)$, where $R = \sqrt{\max_{\mathbf{x} \in X} V_\omega(\mathbf{x}, \mathbf{x}_1)}$ and $B := \sup_{\mathbf{x} \in X} \frac{|f(\mathbf{x}) - f(\mathbf{y})|}{\|\mathbf{x} - \mathbf{y}\|}$.

- The constant may be different. Example of Simplex of $X = \left\{x \in \mathbb{R}^d : x_i \geq 0, \sum_{i=1}^d x_i = 1\right\}$, assume $\|\mathbf{g}\|_\infty \leq 1$,
  - In normal case of $\omega(x) = \frac{1}{2}\|x\|_2^2$, $R^2 \leq 2 = O(1)$, $B \sim \|g\|_2 = O(\sqrt{d})$, overall it's $O(\frac{\sqrt{d}}{\sqrt{T}})$.
  - If we choose $w(x) = \sum_{i=1}^d x_i \ln x_i$ and starting point to be $x_1 = \mathbf{argmin}_{x \in X}\, \omega(x)$, then $\Omega \leq \mathbf{max}_{x\in X}\, \omega(x) - \mathbf{min}_{x \in X}\, \omega(x) = 0 - (-\ln d) = \ln(d)$, overall it's $O(\frac{\ln d}{\sqrt{T}})$
  - The ratio of efficiency is then $O\left(\frac{1}{\ln(d)} \cdot \frac{\mathbf{max}_{x\in X}\|g(x)\|_2}{\mathbf{max}_{x\in X}\|g(x)\|_\infty}\right)$, since $\|g(x)\|_\infty \leq \|g(x)\|_2 \leq \sqrt{d}\|g(x)\|_\infty$, in worst case Mirror Descent is $O(\sqrt{d})$ faster than norm subgradient descent.

Mirror Descent under Smoothness (Exercise 47)

- Lemma H If f convex and gradient Lipschitz w.r.t. some norm, $\|\nabla f(x) - \nabla f(y)\|_{a*} \leq L\|x - y\|_a$, then setting $\gamma_t = 1/L$ will give $\mathbf{min}_{1\leq t\leq T} f(x_t) - f^* \leq L \cdot V_\omega(x^*, x_1)/T$
  - Proof
    - By equivalence of smoothness and Lipschitz, $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) + \mathbf{g}_t^\top(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_a^2$
      - Using the fact of ω is 1-strongly convex, we have $\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_a^2/2 \leq V_\omega(\mathbf{x}_{t+1}, \mathbf{x}_t)$
      - This makes $\frac{1}{\gamma_t}V_\omega(\mathbf{x}_{t+1}, \mathbf{x}_t) + \langle \mathbf{g}_t, \mathbf{x}_{t+1} - \mathbf{x}_t\rangle \geq f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t)$
      - By the optimal condition of prox-mapping, we can show that $V_\omega(\mathbf{x}_{t+1}, \mathbf{x}_t) + \langle \gamma_t \mathbf{g}_t, \mathbf{x}_{t+1}\rangle \leq \underbrace{V_\omega(\mathbf{x}_t, \mathbf{x}_t)}_{=0} + \langle \gamma_t \mathbf{g}_t, \mathbf{x}_t\rangle$
      - This means $f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq 0$, always non-increasing.
    - In the third step of descent lemma, bu convexity we have $V_\omega(\mathbf{x}^\star, \mathbf{x}_t) - V_\omega(\mathbf{x}^\star, \mathbf{x}_{t+1}) + \gamma_t(f(\mathbf{x}^\star) - f(\mathbf{x}_t)) \geq V_\omega(\mathbf{x}_{t+1}, \mathbf{x}_t) + \langle \gamma_t \mathbf{g}_t, \mathbf{x}_{t+1} - \mathbf{x}_t\rangle$
      - take the above into it we get $V_\omega(\mathbf{x}^\star, \mathbf{x}_t) - V_\omega(\mathbf{x}^\star, \mathbf{x}_{t+1}) + \gamma_t(f(\mathbf{x}^\star) - f(\mathbf{x}_t)) \geq \gamma_t(f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t))$
    - take $\gamma_t = L^{-1}$ and sum over $t = [0 : T-1]$, we have
      - $\sum_{i=1}^T f(\mathbf{x}_i) - f(\mathbf{x}^\star) \leq L\left(V_\omega(\mathbf{x}^\star, \mathbf{x}_1) - V_\omega(\mathbf{x}^\star, \mathbf{x}_T)\right)$, QED