Chapter 7 Smoothing Techniques

Convex Conjugate

- Definition 7.1 (Convex conjugate) For any function $f : \mathbb{R}^n \to \mathbb{R}$, its *convex conjugate* is $f^\star(y) = \sup_{x \in \mathbf{dom}(f)} \left\{x^\top y - f(x)\right\}$
- Fenchel's inequality $f^\star(y) \geq x^\top y - f(x), \forall x, y \;\Rightarrow\; x^\top y \leq f(x) + f^*(y), \forall x, y$.
  - which is a generalization if Yong's inequality $x^\top y \leq \frac{\|x\|^2}{2} + \frac{\|y\|_*^2}{2}, \forall x, y$
- Lemma 7.2 (Sec.12, [Roc97], no proof here) If f is convex, lower semi-continuous and proper, then $(f^\star)^\star = f$.
  - *Lower semi-continuous (l.s.c)* means $f(\mathbf{x}) \leq \liminf_{t \to \infty} f\left(\mathbf{x}_t\right)$ for $\mathbf{x}_t \to \mathbf{x}$. *Proper* means $f(x) \geq -\infty$.
- Proposition 7.3 If f is μ-strongly convex then $f^\star$ is continuously differentiable and μ^{-1}-Lipschitz smooth.
  - Proof
    - *Differentiable*
      - Since f strongly-convex -> "f + linear term" strictly convex -> $\sup_{x \in \mathbf{dom}(f)} \left\{x^\top y - f(x)\right\}$ have a unique solution for all $y \in \mathbb{R}^n$, $\mathbf{dom}(f^\star)$ is then convex.
      - $f^\star$ is convex since
        $\sup_{x \in \mathbf{dom}(f)} \left\{x^\top (\lambda y_1 + (1-\lambda) y_2) - f(x)\right\} \geq \lambda \sup_{x \in \mathbf{dom}(f)} \left\{x^\top y_1 - f(x)\right\} + (1-\lambda) \sup_{x \in \mathbf{dom}(f)} \left\{x^\top y_2 - f(x)\right\}$
        - then $\partial f^\star$ is non-empty.
      - Since f strongly convex, its level set $\{x : f(x) \leq \alpha\}$ is closed, and its proper $f(x) > -\infty$. By lemma 7.2, $(f^\star)^\star = f$.
      - Consider $f^\star$'s subdifferential, $g \in \partial f^\star(z) \Leftrightarrow \forall y, f^\star(y) \geq f^\star(z) + g^\top (y - z) \Leftrightarrow \forall y, g^\top y - f^\star(y) \leq g^\top z - f^\star(z)$
        - this means $g^\top z - f^\star(z) = (f^\star)^\star(g) = f(g)$, equivalently,
          $f^\star(z) = g^\top z - f(g) \Leftrightarrow g \in \arg\max_{x \in \mathrm{dom}\, f} \left\{y^T x - f(x)\right\}$.
        - Since f strongly convex, $\arg\max$ is unique, this is a single-valued mapping.
      - By Lemma 26.1 in [Roc97], *∂f is single values mapping iff f essentially smooth. In that case ∂f reduce to gradient mapping ∇f.*
        - This means $f^\star$ is differentiable.
    - *μ^{-1}-smoothness*
      - By similar argument, we know $y \in \partial f(g_y)$, by strong convexity $f(g_z) \geq f(g_y) + y^\top (g_z - g_y) + \mu \|g_y - g_z\|_a^2 / 2$
        - and also $f(g_y) \geq f(g_z) + z^\top (g_y - g_z) + \mu \|g_y - g_z\|_a^2 / 2$
      - combine two, we have $\mu \|g_y - g_z\|_a^2 \leq (g_z - g_y)^\top (z - y) \leq \|g_z - g_y\|_a \|z - y\|_{a*} \Rightarrow \|g_y - g_z\|_a \leq \mu^{-1} \|z - y\|_{a*}$
- Lemma 7.4 (Exercise 49) Let f and g be two proper, convex and semi-continuous functions, then
  - (a) $(f + g)^\star(x) = \inf_y \{f^\star(y) + g^\star(x - y)\}$
  - (b) $(\alpha f)^\star(x) = \alpha f^\star\left(\frac{x}{\alpha}\right)$ for $\alpha > 0$.
  - Proof
    - (a)
      - $(f + g)^\star(x) = \sup_{y \in \mathbf{dom}(f \cap g)} \left\{x^\top y - f(y) - g(y)\right\} = \sup_{y \in \mathbf{dom}(f \cap g)} \inf_z \{(x - z)^\top y - g(y) + f^\star(z)\}$
      - Since function $h(y, z) := (x - z)^\top y - g(y) + f^\star(z)$ is convex w.r.t z and concave w.r.t. y, then by *von Neumann-Fan minimax theorem* $\min \max = \max \min$
        - then $(f + g)^\star(x) = \inf_z \sup_{y \in \mathbf{dom}(f \cap g)} \{(x - z)^\top y - g(y) + f^\star(z)\} = \inf_z \{g^\star(x - z) + f^\star(z)\}$
    - (b) $(\alpha f)^\star(x) = \sup_{y \in \mathbf{dom}(f \cap g)} \left\{x^\top y - \alpha f(y)\right\} = \alpha \sup_{y \in \mathbf{dom}(f \cap g)} \left\{\alpha^{-1} x^\top y - f(y)\right\} = \alpha f^\star(x / \alpha)$.
- Examples
  - $f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^\top Q \mathbf{x}, Q \succ 0, f^\star(y) = \frac{1}{2} \mathbf{y}^\top Q^{-1} \mathbf{y}$.
  - $f(\mathbf{x}) = \sum_{i=1}^n x_i \log(x_i), f^\star(\mathbf{y}) = \sum_{i=1}^n e^{y_i - 1}$.
  - $f(\mathbf{x}) = -\sum_{i=1}^n \log(x_i), f^\star(\mathbf{y}) = -\sum_{i=1}^n \log(-y_i) - n$.
  - $f(\mathbf{x}) = \|\mathbf{x}\|, f^\star(\mathbf{y}) = \begin{cases} 0, & \|\mathbf{y}\|_* \leq 1 \\ +\infty, & \|\mathbf{y}\|_* > 1. \end{cases}$

Smoothing Techniques

Common Smoothing techs

- Nesterov's smoothing based on conjugate function $f_\mu(x) = \max_{y \in \mathbf{dom}(f^*)} \left\{x^\top y - f^\star(y) - \mu \cdot d(y)\right\} = (f^\star + \mu d)^\star(x)$, where $d(y)$ is some proximity function, d is (i) 1-strongly convex (ii) non-negative everywhere and (possibly iii) $\min d(y) = 0$.
  - By Prop 7.3, f_μ is *continuously differentiable* and μ^{-1}-Lipschitz.
  - Examples of d:

    - $d(\mathbf{y}) = \frac{1}{2}\|\mathbf{y}-\mathbf{y}_0\|_2^2$
    - $d(\mathbf{y}) = \frac{1}{2}\sum w_i(y_i - y_{0,i})^2$ with $w_i \geq 1$
    - $d(\mathbf{y}) = V_\omega(\mathbf{y}, \mathbf{y}_0)$.
- Moreau-Yosida smoothing/regularization $f_\mu(x) = \min_{y\in\text{dom}(f)}\left\{f(y) + \frac{1}{2\mu}\|x-y\|_2^2\right\}$.
    - when $d(y) = \frac{1}{2}\|y\|_2^2$, Nesterov's smoothing turns to Moreau-Yosida regularization (by Lemma 7.4).
- Lasry-Lions regularization double application of M-Y reg $f_{\mu,\delta}(x) = \max_y \min_z \left\{f(z) + \frac{1}{2\mu}\|z-y\|_2^2 - \frac{1}{2\delta}\|y-x\|_2^2\right\}$
    - If f 1-Lipschitz, then choosing $(\delta, \mu) = O(\epsilon)$ guarantees ϵ approximation error, and $f_{\mu,\delta}(\mathbf{x})$ is $O(1/\epsilon)$-smooth.
    - Computation inefficiency due to solving nonconvex problems.
- Randomized smoothing $f_\mu(x) = \mathbb{E}_Z f(x+\mu Z)$ where Z is isotopic Gaussian or uniform.
    - Choosing $\mu = O(\epsilon)$ guarantees ϵ☒ approximation error.
    - $f_\mu(\mathbf{x})$ is $O\left(\frac{\sqrt{d}}{\epsilon}\right)$-smooth, dimension dependent.
    - Stochastic gradient is computationally efficient through sampling.
- Ben-Tal-Teboulle smoothing based on recession function. Only applicable to particular class $f(x) = F\left(f_1(x), f_2(x), \ldots, f_m(x)\right)$, where $F(y) = \max_{x\in\text{dom}(g)}\{g(x+y) - g(x)\}$ is the recession function of some func g.
    - The smoothed function is $f_\mu(x) = \mu g\left(\frac{f_1(x)}{\mu}, \ldots, \frac{f_m(x)}{\mu}\right)$.

Nesterov's Smoothing

- Proposition 7.10 $\forall \mu > 0$, let $D_Y^2 = \max_{y\in Y} d(y)$, $f(x) - \mu D_Y^2 \leq f_\mu(x) \leq f(x)$.
    - Proof
        - Left: $f_\mu(x) = \max_{y\in\text{dom}(f^*)}\left\{x^\top y - f^\star(y) - \mu\cdot d(y)\right\} \geq \max\left\{x^\top y - f^\star(y)\right\} - \mu\max\left\{d(y)\right\} = f(x) - \mu D^2$
        - Right: $f_\mu(x) = \max_{y\in\text{dom}(f^*)}\left\{x^\top y - f^\star(y) - \mu\cdot d(y)\right\} \leq f_\mu(x) = \max_{y\in\text{dom}(f^*)}\left\{x^\top y - f^\star(y)\right\} = f(x)$
- Remark Tadeoff b.t.w. approximation error and optimization efficiency $f(\mathbf{x}) - f^* \leq \underbrace{f(\mathbf{x}) - f_\mu(\mathbf{x})}_{\text{approximation error}} + \underbrace{f_\mu(\mathbf{x}) - \min_x f_\mu(\mathbf{x})}_{\text{optimization error}}$.
- Suppose we can access $\nabla f_\mu(x)$ and apply gradient methods to solve $\min_{x\in X} f_\mu(x)$
    - PGD gives $f\left(x_t\right) - f_* \leq O\left(\frac{R^2}{\mu t} + \mu D_Y^2\right)$
        - to achieve arror of ϵ, we have to set $\mu = O\left(\frac{\epsilon}{D_Y^2}\right)$, and the number of iteration is $T = O\left(\frac{R^2 D_Y^2}{\epsilon^2}\right)$
    - Accelerater GD gives $f\left(x_t\right) - f_* \leq O\left(\frac{R^2}{\mu t^2} + \mu D_Y^2\right)$
        - similarly to achieve error of ϵ, we have to set $\mu = O\left(\frac{\epsilon}{D_Y^2}\right)$ and $T = O\left(\frac{RD_Y}{\epsilon}\right)$

Example

- The function form is generalized as $f(x) = \max_{y\in Y}\{\langle Ax+b, y\rangle - \phi(y)\}$, where ϕ convex and continuous, Y convex and compact.
    - example like $f(x) = \max_{1\leq i\leq m}|a_i^T x - b_i|$ can be re-written as $f(x) = \max_{y\in Y}\sum_{i=1}^m \left(a_i^T x - b_i\right)y_i$, where $Y := \left\{y\in\mathbb{R}^m : \sum_{i=1}^m |y_i| \leq 1\right\}$.
    - The smoothed function is $f_\mu(x) = \max_{y\in Y}\{\langle Ax+b, y\rangle - \phi(y) - \mu\cdot d(y)\} = (\phi + \mu\cdot d)^\star(Ax+b)$
- As an example, $f(x) = |x| = \sup_{|y|\leq 1} yx = \sup_{\substack{y_1,y_2\geq 0\\ y_1+y_2=1}} (y_1 - y_2)x$,
    - $d(y) = y^2/2$, then $f_\mu(x) = \sup_{|y|\leq 1}\left\{yx - \frac{\mu}{2}y^2\right\} = \begin{cases}\frac{x^2}{2\mu}, & |x|\leq\mu\\ |x| - \frac{\mu}{2}, & |x| > \mu\end{cases}$.
    - $d(y) = 1 - \sqrt{1-y^2}$ (half circle), clearly 1-strongly convex. $f_\mu(x) = \sup_{|y|\leq 1}\left\{yx - \mu\left(1 - \sqrt{1-y^2}\right)\right\} = \sqrt{x^2+\mu^2} - \mu$.
        - This is a special case of Ben-Tal & Teboulle's smoothing $|x| = \sup_y\{g(x+\mu) - g(y)\}, g(y) = \sqrt{1+y^2}$ and $f_\mu(x) = \mu g\left(x/\mu\right) = \sqrt{x^2+\mu^2}$
    - $d(y) = y_1\log y_1 + y_2\log y_2 + \log 2$, where $Y = \{(y_1, y_2) : y_1, y_2 \geq 0, y_1 + y_2 = 1\}$, $f_\mu(x) = \mu\log\left(\frac{e^{-\frac{x}{\mu}} + e^{\frac{x}{\mu}}}{2}\right)$
        - This can be seen as $|x| = \max\{x, -x\} = \sup_y\{g(x+\mu) - g(y)\}, \quad g(y) = \log\left(e^{y_1} + e^{y_2}\right)$.

Moreau-Yosida Regularization

- Definition 7.11 the *proximal operator* of convex f at a given point x is defined as $\text{prox}_f(x) = \underset{y}{\text{argmin}}\left\{f(y) + \frac{1}{2}\|x-y\|^2\right\}$.
- Examples
    - Indicator function $f(\mathbf{x}) = \delta_X(\mathbf{x}) = \begin{cases}0, & \mathbf{x}\in X\\ +\infty, & \mathbf{x}\notin X\end{cases}$, $\text{prox}_f(\mathbf{x}) = \Pi_X(\mathbf{x})$ is the projection.

- $f(\mathbf{x}) = \mu\|\mathbf{x}\|_1$, $\mathbf{prox}_{\mu|\cdot|}(x_i) = \begin{cases} x_i - \mu & \text{if } x_i > \mu \\ 0 & \text{if } |x_i| \leq \mu \\ x_i + \mu & \text{if } x_i < -\mu \end{cases} = \mathbf{sign}(\mathbf{x}) \otimes [|\mathbf{x}| - \lambda] +$ *soft thresholding operator*.
- 2-norm $f(\mathbf{x}) := \|\mathbf{x}\|_2$, $\mathbf{prox}_{\lambda f}(\mathbf{x}) = [1 - \lambda/\|\mathbf{x}\|_2] + \mathbf{x}$
- Support function $f(\mathbf{x}) := \mathbf{max}_{\mathbf{y} \in \mathcal{C}}\, \mathbf{x}^T \mathbf{y}$, $\mathbf{prox}_{\lambda_f}(\mathbf{x}) = \mathbf{x} - \lambda \pi_{\mathcal{C}}(\mathbf{x})$
- Square norm $f(\mathbf{x}) := (1/2)\|\mathbf{x}\|_2^2$, $\mathbf{prox}_{\lambda f}(\mathbf{x}) = (1/(1+\lambda))\mathbf{x}$
- log function $f(\mathbf{x}) := -\log(x)$, $\mathbf{prox}_{\lambda f}(x) = \left(\left(x^2 + 4\lambda\right)^{1/2} + x\right)/2$
- log det $f(\mathbf{x}) := -\log\det(\mathbf{X})$, $\mathbf{prox}_{\lambda f}(x)$ is the log function prox applied to the individual eigenvalues of $\mathbf{X}$.

- **Proposition 7.13** f convex, then
  - (a) *fixed point* $x^\star$ minimized $f(x)$ iff $x^\star = \mathbf{prox}_f(x^\star)$
    - **Proof** (one direction is obvious)
      - $x^\star = \mathbf{prox}_f(x^\star) \Leftrightarrow \forall y, f(x^\star) \leq f(y) + \frac{1}{2}\|x^\star - y\|^2$
      - If $\exists y_0$ s.t. $f(y) < f(x^\star)$, by convexity,
        $f(y_\lambda) := f(x^\star + \lambda(y - x^\star)) \leq f(x^\star) - \lambda(f(x^\star) - f(y)) \Rightarrow f(x^\star) \geq f(x^\star + \lambda(y - x^\star)) + \lambda(f(x^\star) - f(y))$
      - Setting $\lambda \mathbf{min}\{1, \alpha 2(f(x^\star) - f(y))/\|x^\star - y\|^2\}$, we get $f(x^\star) \geq f(y_\lambda) + \frac{\alpha}{2}\|x^\star - y_\lambda\|^2$, setting $\alpha > 1$ and we get contradiction.
      - Another path is to prove $0 \in \partial f(x_*) + (x_* - x_*) \Longrightarrow 0 \in \partial f(x_*)$.
  - (b) *Non-expansive* $\|\mathbf{prox}_f(x) - \mathbf{prox}_f(y)\| \leq \|x - y\|$
    - **Proof**
      - by optimial condition (suppose $\mathbf{dom}(f) = \mathbb{R}^n$), $\nabla f(\mathbf{prox}(x)) + \mathbf{prox}(x) - x = 0$ and $\nabla f(\mathbf{prox}(y)) + \mathbf{prox}(y) - y = 0$
      - By convexity $(\nabla f(\mathbf{prox}(y)) - \nabla f(\mathbf{prox}(x)))^\top (\mathbf{prox}(y) - \mathbf{prox}(x)) \geq 0$
        - This means $(y - x)^\top (\mathbf{prox}(y) - \mathbf{prox}(x)) \geq \|\mathbf{prox}(y) - \mathbf{prox}(x)\|_2^2 \Rightarrow$ QED
  - (c) *Moreau Decomposition* $x = \mathbf{prox}_f(x) + \mathbf{prox}_{f^\star}(x)$
    - **Proof**
      - By fact of $\partial f^\star(g) = \mathbf{arg\,max}_x\{g^\top x - f(x)\} = \{x | \nabla f(x) = g\}$
      - we already have $\nabla f(\mathbf{prox}(x)) + \mathbf{prox}(x) - x = 0$
        - which can be seen as $g = \nabla f(\mathbf{prox}(x))$, so $\nabla f^\star(g) = \mathbf{prox}(x))$
        - then this means $g + \nabla f^\star(g) - x = 0$, which is the optimial condition for $\mathbf{prox}_{f^\star}(x)$, QED.

Proximal Point Algorithm

- **Proximal Point Algorithm (PPA)** $\mathbf{x}_{t+1} = \mathbf{prox}_{\gamma t f}(\mathbf{x}_t)$
  - This comes from $x_{t+1} = x_t - L^{-1}\nabla f_\mu(x_t)$, (Danskin's thm used here)
- **Example** $\mathbf{min}_x \|Ax + b\|_2^2 + \lambda\|x\|_1 = \mathbf{min}_x\, f_1 + f_2$, where f_1 differentiable while f_2 is not. Usually we approximate $x_{t+1} = \mathbf{prox}_{f_2}(x_t - \frac{1}{L_{f_1}}\nabla f_1(x_t))$.
- **Theorem 7.14** If f convex, then PPA satisfies $f(x_t) - f^\star \leq \frac{\|x_0 - x_*\|_2^2}{2\sum_{\tau=0}^{t-1}\gamma_\tau}$
  - **Proof**
    - by the optimality condition of x_{t+1}, $f(x_{t+1}) + \frac{1}{2\gamma_t}\|x_{t+1} - x_t\|_2^2 \leq f(x_t)$, this is something like sufficient decrease.
    - by first order optimality condition, $0 \in \partial f(x_{t+1}) + \frac{1}{\gamma_t}(x_{t+1} - x_t) \Longrightarrow \frac{x_t - x_{t+1}}{\gamma_t} \in \partial f(x_{t+1})$,
      - so by convexity/subgradient
        $f(x_{\tau+1}) - f^\star \leq \frac{1}{\gamma_\tau}(x_\tau - x_{\tau+1})^T(x_{\tau+1} - x_*) = \frac{1}{\gamma_t}\left[(x_\tau - x_*)^T(x_{\tau+1} - x_*) - \|x_{\tau+1} - x_*\|^2\right]$
      - By Young's inequality, $(x_\tau - x_*)^T(x_{\tau+1} - x_*) \leq \frac{1}{2}\left[\|x_\tau - x_*\|^2 + \|x_{\tau+1} - x_*\|^2\right]$, plug this in and we get
      - $\gamma_\tau(f(x_{\tau+1}) - f^\star) \leq \frac{1}{2}\left[\|x_\tau - x_*\|^2 - \|x_{\tau+1} - x_*\|^2\right]$
      - Then we can happily do our summation over $t \in [0 : T-1]$
    - Since we have monotonicity in f, this algorithm have last iteration convergence.
  - *Comment* prox step is more like a second order step.... more computation, but also more effective.
  - Algorithm converge a.l.a $\sum_t \gamma_t \to \infty$.
  - Larger γ_t makes algo converge faster, but harder to solve each step.