Google Research

Duality in machine learning

Mathieu Blondel

November 26, 2020

Outline

Closed functions

- The domain of a function is denoted $\mathrm{dom}(f) = \{x \in \mathbb{R}^d \colon f(x) < \infty\}$
- A function is closed if for all $\alpha \in \mathbb{R}$ the sub-level set

$$\{x \in \mathrm{dom}(f) \colon f(x) \leq \alpha\}$$

  is closed (reminder: a set is closed if it contains its boundary)
- If f is continuous and $\mathrm{dom}(f)$ is closed then f is closed
- Example 1: $f(x) = x \log x$ is not closed over $\mathrm{dom}(f) = \mathbb{R}_{>0}$
- Example 2: $f(x) = x \log x$ is closed over $\mathrm{dom}(f) = \mathbb{R}_{\geq 0}$, $f(0) = 0$
- Example 3: the indicator function $I_{\mathcal{C}}$ is closed if $\mathcal{C}$ is closed

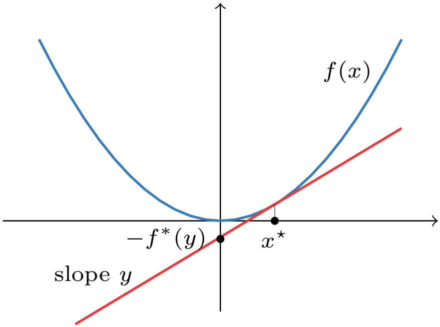
$$I_{\mathcal{C}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{C} \\ \infty & \text{if } x \notin \mathcal{C}. \end{cases}$$

Convex conjugate

- Fix a slope y. What is the intercept b of the tighest linear lower bound of f? In other words, for all $x \in \mathrm{dom}(f)$, we want

$$\begin{aligned} \langle x, y \rangle - b \leq f(x) &\Leftrightarrow \langle x, y \rangle - f(x) \leq b \\ &\Leftrightarrow b = \sup_{x \in \mathrm{dom}(f)} \langle x, y \rangle - f(x) \end{aligned}$$

- The value of the intercept is denoted $f^*(y)$, the conjugate of $f(x)$.



Convex conjugate

- Equivalent definition

$$-f^*(y) = \inf_{x \in \mathrm{dom}(f)} f(x) - \langle x, y \rangle$$

- f^* can take values on the extended real line $\mathbb{R} \cup \{\infty\}$
- f^* is closed and convex (even when f is not)
- Fenchel-Young inequality: for all x, y

$$f(x) + f^*(y) \geq \langle x, y \rangle$$

Convex conjugate examples

- Example 1: $f(x) = I_{\mathcal{C}}(x)$, the indicator function of $\mathcal{C}$

$$f^*(y) = \sup_{x \in \mathrm{dom}(f)} \langle x, y \rangle - f(x) = \sup_{x \in \mathcal{C}} \langle x, y \rangle$$

f^* is called the support function of $\mathcal{C}$

- Example 2: $f(x) = \langle x, \log x \rangle$, then

$$f^*(y) = \sum_{i=1}^{d} e^{y_i - 1}$$

- Example 3: $f(x) = \langle x, \log x \rangle + I_{\triangle^d}(x)$

$$f^*(y) = \frac{\exp(y)}{\sum_{i=1}^{d} \exp(y_i)}$$

Convex conjugate calculus

- Separable sum

$$f(x) = \sum_{i=1}^{d} f_i(x_i) \qquad f^*(y) = \sum_{i=1}^{d} f_i^*(y_i)$$

- Scalar multiplication ($c > 0$)

$$f(x) = c \cdot g(x) \qquad f^*(y) = c \cdot g^*(y/c)$$

- Addition to affine function / translation of argument

$$f(x) = g(x) + \langle a, x \rangle + b \qquad f^*(y) = g^*(y - a) - b$$

- Composition with invertible linear mapping

$$f(x) = g(Ax) \qquad f^*(y) = g^*(A^{-T} y)$$

Biconjugates

- The bi-conjugate

$$f^{**}(x) = \sup_{y \in \mathrm{dom}(f^*)} \langle x, y \rangle - f^*(y)$$

- f^{**} is closed and convex
- If f is closed and convex then

$$f^{**}(x) = f(x)$$

- If f is not convex, f^{**} is the tightest convex lower bound of f

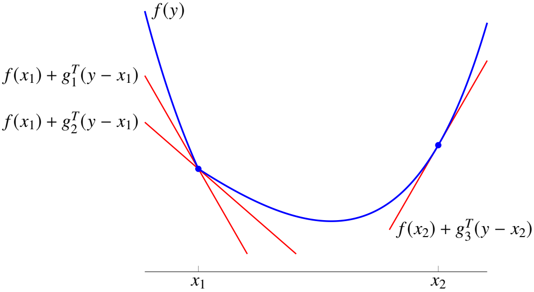
Subgradients

- Recall that a differentiable convex function always lies above its tangents, i.e., for all $x, y \in \mathrm{dom}(f)$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

- g is the subgradient of a convex function f if for all $x, y \in \mathrm{dom}(f)$

$$f(y) \geq f(x) + \langle g, y - x \rangle$$

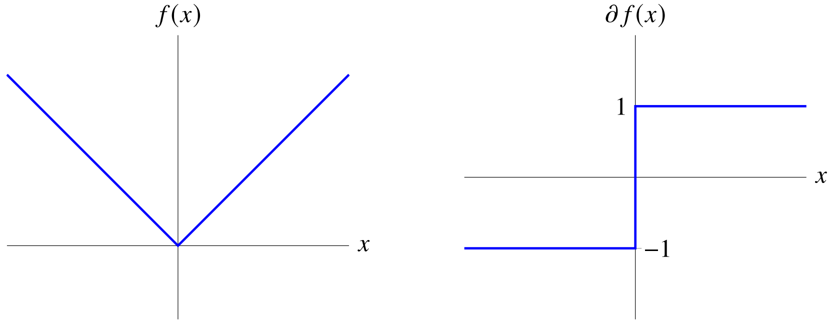


Subdifferential

- The subdifferential is the set of all subgradients

$$\partial f(x) = \{g\colon f(y) \geq f(x) + \langle g, y - x \rangle \ \forall y \in \mathrm{dom}(f)\}$$

- Example: $f(x) = |x|$



$$\partial f(0) = [-1, 1] \qquad \partial f(x) = \{\nabla f(x)\} \text{ if } x \neq 0$$

Conjugates and subdifferentials

- Alternative definition of subdifferential

$$\partial f^*(y) = \{x \in \mathrm{dom}(f) \colon f(x) + f^*(y) = \langle x, y \rangle\}$$

- From Danskin's theorem

$$\partial f^*(y) = \operatorname*{argmax}_{x \in \mathrm{dom}(f)} \langle x, y \rangle - f(x)$$

- If f is strictly convex

$$\nabla f^*(y) = \operatorname*{argmax}_{x \in \mathrm{dom}(f)} \langle x, y \rangle - f(x)$$

- And similarly for $\partial f(x)$, $\nabla f(x)$

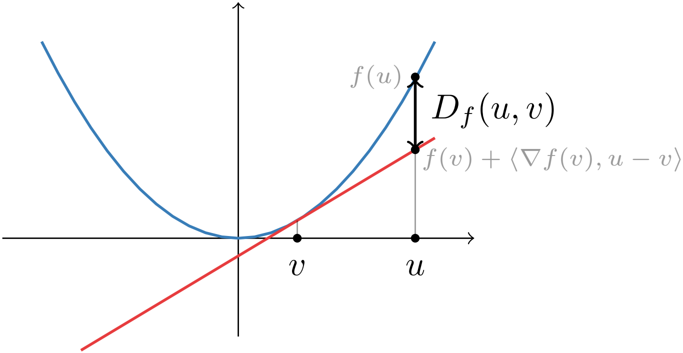
Outline

Bregman divergences

- Let f be convex and differentiable.
- The Bregman divergence generated by f between u and v is

$$D_f(u, v) = f(u) - f(v) - \langle \nabla f(v), u - v \rangle$$

- It is the difference between $f(u)$ and its linearization around v.



Bregman divergences

- Recall that a differentiable convex function always lies above its tangents, i.e., for all u, v

$$f(u) \geq f(v) + \langle \nabla f(v), u - v \rangle$$

- The Bregman divergence is thus non-negative for all u, v

$$D_f(u, v) \geq 0$$

- Put differently, a differentiable function f is convex if and only if it generates a non-negative Bregman divergence.

- Not necessarily symmetric

Bregman divergences

- Example 1: if $f(x) = \frac{1}{2}\|x\|_2^2$, then D_f is the squared Euclidean distance

$$D_f(u, v) = \frac{1}{2}\|u - v\|_2^2$$

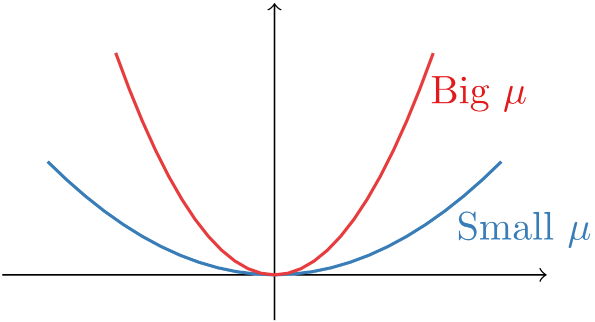
- Example 2: if $f(x) = \langle x, \log x \rangle$, then D_f is the (generalized) Kullback-Leibler divergence

$$D_f(p, q) = \sum_{i=1}^{d} p_i \log \frac{p_i}{q_i} - \sum_{i=1}^{d} p_i + \sum_{i=1}^{d} q_i$$

Strong convexity

- f is said to be μ-strongly convex w.r.t. a norm $\|\cdot\|$ over $\mathcal{C}$ if

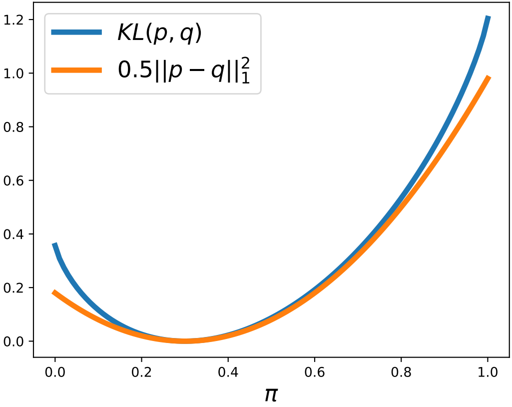
$$\frac{\mu}{2}\|u - v\|^2 \leq D_f(u, v) \quad \text{for all} \quad u, v \in \mathcal{C}$$



- Example 1: $f(x) = \frac{1}{2}\|x\|_2^2$ is 1-strongly convex w.r.t. $\|\cdot\|_2$ over $\mathbb{R}^d$.
- Example 2: $f(x) = \langle x, \log x \rangle$ is 1-strongly convex w.r.t. $\|\cdot\|_1$ over the probability simplex $\triangle^d = \{p \in \mathbb{R}_+^d : \|p\|_1 = 1\}$.

Strong convexity

Pinsker's inequality

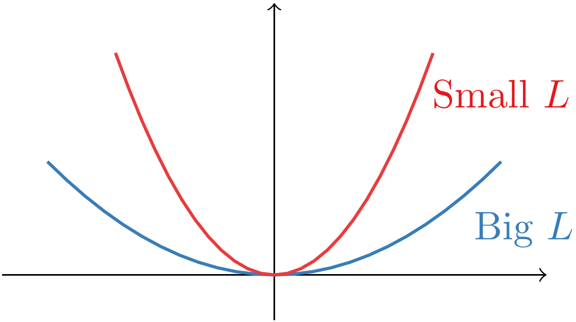


$$p = (\pi, 1-\pi), q = (0.3, 0.7)$$

Smoothness

- f is said to be L-smooth w.r.t. a norm $\|\cdot\|$ over $\mathcal{C}$ if

$$D_f(u, v) \leq \frac{L}{2}\|u - v\|^2 \quad \text{for all} \quad u, v \in \mathcal{C}$$



- Example 1: $f(x) = \frac{1}{2}\|x\|_2^2$ is 1-smooth w.r.t. $\|\cdot\|_2$ over $\mathbb{R}^d$.
- Example 2: $f(x) = \log \sum_i e^{x_i}$ is 1-smooth w.r.t. $\|\cdot\|_\infty$ over $\mathbb{R}^d$

Hessian bounds

- When f is twice differentiable, this also leads to bounds on $\nabla^2 f$
- When f is strongly convex, we have

$$\mu \cdot \mathrm{Id}_d \preceq \nabla^2 f$$

- When f is smooth, we have

$$\nabla^2 f \preceq L \cdot \mathrm{Id}_d$$

- Functions can be both strongly-convex and smooth, e.g., the sum of a smooth function and a strongly-convex function.

Lipschitz functions

- Given a norm $\|x\|$ on $\mathcal{C}$, its dual (also on $\mathcal{C}$) is

$$\|y\|_* = \max_{\|x\| \leq 1} \langle x, y \rangle$$

  Examples: $|| \cdot \|_2$ is dual with itself, $\| \cdot \|_1$ is dual with $\| \cdot \|_\infty$

- A function $g \colon \mathbb{R}^d \to \mathbb{R}^p$ is said to be L-Lipschitz continuous w.r.t. $\| \cdot \|$ over $\mathcal{C}$ if for all $x, y \in \mathcal{C} \subseteq \mathbb{R}^d$

$$\|g(x) - g(y)\|_* \leq L\|x - y\|$$

- Choose $g = \nabla f$. Then f is said to have Lipschitz-continuous gradients.

- **Fact.** A function is L-smooth if and only if it has L-Lipschitz continuous gradients.

Strong convexity and smoothness duality

- **Theorem.**

  f is μ-strongly convex w.r.t. $\|\cdot\| \Leftrightarrow f^*$ is $\frac{1}{\mu}$-smooth w.r.t. $\|\cdot\|_*$

- Example 1:
  $f(x) = \frac{\mu}{2}\|x\|^2$ is μ-strongly convex w.r.t. $\|\cdot\|$,
  $f^*(y) = \frac{1}{2\mu}\|y\|_*^2$ is $\frac{1}{\mu}$-smooth w.r.t. $\|\cdot\|_*$.

- Example 2:
  $f(x) = \langle x, \log x\rangle$ is 1-strongly convex w.r.t. $\|\cdot\|_1$ over $\triangle^d$,
  $f^*(y) = \log \sum_i e^{y_i}$ is 1-smooth w.r.t. $\|\cdot\|_\infty$ over $\mathbb{R}^d$.

Smoothing: Moreau-Yosida regularization

- Suppose we have a non-smooth function $g(x)$, e.g., $g(x) = |x|$
- We can create a smooth version of g by

$$g_\mu(x) = \min_u g(u) + \frac{1}{2\mu}\|x - u\|_2^2$$

- This is also called the inf-convolution of g with $\frac{1}{2\mu}\|\cdot\|_2^2$
- The gradient of g_μ is equal to the proximity operator of μg

$$\begin{aligned}\nabla g_\mu(x) &= u^\star \\ &= \operatorname*{argmin}_u g(u) + \frac{1}{2\mu}\|x - u\|_2^2 \\ &= \operatorname*{argmin}_u \mu g(u) + \frac{1}{2}\|x - u\|_2^2 \\ &= \mathrm{prox}_{\mu g}(x)\end{aligned}$$

Smoothing: Moreau-Yosida regularization

- Example: $g(x) = |x|$
- The proximity operator is the **soft-thresholding** operator

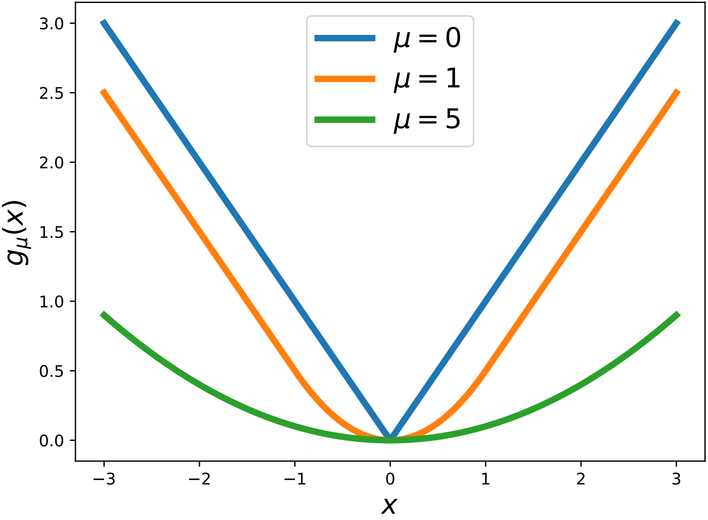
$$u^\star = \operatorname*{argmin}_u \mu|u| + \frac{1}{2}\|x - u\|_2^2 = \begin{cases} 0 & \text{if } |x| \le \mu \\ x - \mu \text{sign}(x) & \text{if } |x| > \mu. \end{cases}$$

- Using $g_\mu(x) = |u^\star| + \frac{1}{2\mu}\|x - u^\star\|_2^2$, we get

$$g_\mu(x) = \begin{cases} \frac{x^2}{2\mu} & \text{if } |x| \le \mu \\ |x| - \frac{\mu}{2} & \text{if } |x| > \mu. \end{cases}$$

- This is known as the Huber loss.

Smoothing: Moreau-Yosida regularization



Smoothing: dual approach

- Suppose we want to smooth a convex function $g(x)$
- **Step 1:** derive the conjugate $g^*(y)$
- **Step 2:** add regularization

$$g^*_\mu(y) = g^*(y) + \frac{\mu}{2}\|y\|_2^2$$

- **Step 3:** derive the bi-conjugate

$$g^{**}_\mu(x) = g_\mu(x) = \max_{y \in \mathrm{dom}(g^*)} \langle x, y\rangle - g^*_\mu(y)$$

- Equivalent (dual) to Moreau-Yosida regularization!
- By duality, $g_\mu(x)$ is $\frac{1}{\mu}$-smooth since $\frac{\mu}{2}\|\cdot\|_2^2$ is μ-strongly convex.

Smoothing: dual approach

- Example 1: $g(x) = |x|$
- **Step 1:** $g^*(y) = I_{[-1,1]}(y)$
- **Step 2:** add regularization

$$g^*_\mu(y) = I_{[-1,1]}(y) + \frac{\mu}{2} y^2$$

- **Step 3:** derive the bi-conjugate

$$g^{**}_\mu(x) = g_\mu(x) = \max_{y \in [-1,1]} x \cdot y - \frac{\mu}{2} y^2$$

- **Solution:**

$$g_\mu(x) = x \cdot y^\star - \frac{\mu}{2}(y^\star)^2 \quad \text{where} \quad y^\star = \text{clipping}_{[-1,1]}\left(\frac{1}{\mu}x\right)$$

Smoothing: dual approach

- Example 2: $g(x) = \max(0, x)$, i.e., the relu function
- **Step 1:** $g^*(y) = I_{[0,1]}(y)$
- **Step 2:** add regularization

$$g_\mu^*(y) = I_{[0,1]}(y) + \frac{\mu}{2} y^2$$

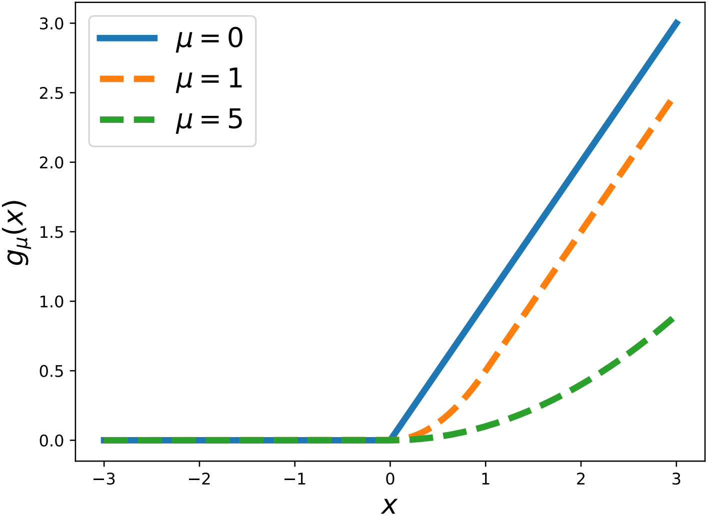
- **Step 3:** derive the bi-conjugate

$$g_\mu^{**}(x) = g_\mu(x) = \max_{y \in [0,1]} x \cdot y - \frac{\mu}{2} y^2$$

- **Solution:**

$$g_\mu(x) = x \cdot y^\star - \frac{\mu}{2} (y^\star)^2 \quad \text{where} \quad y^\star = \text{clipping}_{[0,1]}\left(\frac{1}{\mu} x\right)$$

Smoothing: dual approach



Smoothing: dual approach

- Regularization is not limited to $\frac{\mu}{2}\|y\|^2$
- Any strongly-convex regularization can be used
- Example: softmax

$$
\begin{aligned}
g(x) &= \max_{i \in \{1,\ldots,d\}} x_i \\
g^*(y) &= I_{\triangle^d}(y) \\
g^*_\mu(y) &= I_{\triangle^d}(y) + \mu\langle y, \log y\rangle \\
g_\mu(x) &= \mu \log \sum_{i=1}^{d} \exp(x_i/\mu) \\
\nabla g_\mu(x) &= \frac{\exp(x/\mu)}{\sum_{i=1}^{d} \exp(x_i/\mu)}
\end{aligned}
$$

Outline

Fenchel dual

- $F(\theta)$ convex, $G(W)$ strictly convex
- We are going to derive the Fenchel dual of

$$\min_{W \in \mathbb{R}^{d \times k}} F(XW) + G(W)$$

  where $X \in \mathbb{R}^{n \times d}$ and $W \in \mathbb{R}^{d \times k}$

- Let us rewrite the problem using constraints

$$\min_{\substack{W \in \mathbb{R}^{d \times k} \\ \theta \in \mathbb{R}^{n \times k}}} F(\theta) + G(W) \text{ s.t. } \theta = XW$$

  F and G now involve different variables (tied by equality constraints)

Fenchel dual

- We now use Lagrange duality

$$\min_{\substack{W \in \mathbb{R}^{d \times k} \\ \theta \in \mathbb{R}^{n \times k}}} \max_{\alpha \in \mathbb{R}^{n \times k}} F(\theta) + G(W) + \langle \alpha, \theta - XW \rangle$$

- Since the problem only has linear constraints and is feasible, strong duality holds (we can swap the min and max)

$$\max_{\alpha \in \mathbb{R}^{n \times k}} \min_{\substack{W \in \mathbb{R}^{d \times k} \\ \theta \in \mathbb{R}^{n \times k}}} F(\theta) + G(W) + \langle \alpha, \theta - XW \rangle$$

- We are now going to introduce the convex conjugates of F and G.

Fenchel dual

- For the terms involving θ, we have

$$\min_{\theta \in \mathbb{R}^{n \times k}} F(\theta) + \langle \alpha, \theta \rangle = -F^*(-\alpha)$$

- For the terms involving W, we have

$$\begin{aligned} \min_{W \in \mathbb{R}^{k \times d}} G(W) - \langle \alpha, XW \rangle &= \min_{W \in \mathbb{R}^{d \times k}} G(W) - \langle W, X^\top \alpha \rangle \\ &= -G^*(X^\top \alpha) \end{aligned}$$

- To summarize, the dual consists in solving

$$\max_{\alpha \in \mathbb{R}^{n \times k}} -F^*(-\alpha) - G^*(X^\top \alpha)$$

- The primal-dual link is

$$W^\star = \nabla G^*(X^\top \alpha^\star)$$

Fenchel dual for loss sums

- Typically, in machine learning, F is a sum of loss terms and G is a regularization term:

$$F(\theta) = \sum_{i=1}^{n} L(\theta_i, y_i) \quad \text{where} \quad \theta_i = W^\top x_i$$

- Since the sum is separable, we get

$$F^*(-\alpha) = \sum_{i=1}^{n} L^*(-\alpha_i, y_i)$$

  where L^* is the convex conjugate in the first argument of L

- What have we gained? If G^* is simple enough, we can solve the objective by dual block coordinate ascent.

Examples of regularizer

- Squared L_2 norm

$$G(W) = \frac{\lambda}{2}\|W\|_F^2 = \frac{\lambda}{2}\langle W, W\rangle$$
$$G^*(V) = \frac{1}{2\lambda}\|V\|_F^2$$
$$\nabla G^*(V) = \frac{1}{\lambda}V$$

- Elastic-net

$$G(W) = \frac{\lambda}{2}\|W\|_F^2 + \lambda\rho\|W\|_1$$
$$G^*(V) = \langle \nabla G^*(V), V\rangle - G(\nabla G^*(V))$$
$$\nabla G^*(V) = \operatorname*{argmin}_W \frac{1}{2}\|W - V/\lambda\|_F^2 + \rho\|W\|_1$$

The last operation is the soft-thresholding operator (element-wise).

Fenchel-Young losses

- The Fenchel-Young loss generated by Ω

$$L_\Omega(\theta_i, y_i) = \Omega^*(\theta_i) + \Omega(y_i) - \langle \theta_i, y_i \rangle$$

- Non-negative (Fenchel-Young inequality)
- Convex in θ even when Ω is not
- If Ω is strictly convex, the loss is zero if and only if

$$y_i = \nabla \Omega^*(\theta_i) = \underset{y' \in \mathrm{dom}(\Omega)}{\mathrm{argmax}} \, \langle y', \theta_i \rangle - \Omega(y')$$

- Conjugate function (in the first argument)

$$L_\Omega^*(-\alpha_i, y_i) = \Omega(y_i - \alpha_i) - \Omega(y_i)$$

Fenchel-Young losses

- Squared loss

$$\Omega(\beta_i) = \frac{1}{2}\|\beta_i\|_2^2 \qquad L_\Omega(\theta_i, y_i) = \frac{1}{2}\|y_i - \theta_i\|_2^2$$

$y_i \in \mathbb{R}^k$

- Multiclass perceptron loss

$$\Omega(\beta_i) = I_{\triangle^k}(\beta_i) \qquad L_\Omega(\theta_i, y_i) = \max_{j \in \{1,\ldots,k\}} \theta_{i,j} - \langle \theta_i, y_i \rangle$$

$y_i \in \{e_1, \ldots, e_k\}$

- Multiclass hinge loss

$$\Omega(\beta_i) = I_{\triangle^k}(\beta_i) - \langle \beta_i, v_i \rangle \qquad L_\Omega(\theta_i, y_i) = \max_{j \in \{1,\ldots,k\}} \theta_{i,j} + v_{i,j} - \langle \theta_i, y_i \rangle$$

$v_i = 1 - y_i \qquad y_i \in \{e_1, \ldots, e_k\}$

Dual in the case of Fenchel-Young losses

- Recall that the dual is

$$\max_{\alpha \in \mathbb{R}^{n \times k}} -\sum_{i=1}^{n} L^*(-\alpha_i, y_i) - G^*(X^\top \alpha)$$

  with primal-dual link $W^\star = \nabla G^*(X^\top \alpha^\star)$

- Using the change of variable $\beta_i = y_i - \alpha_i$ and $L = L_\Omega$, we obtain

$$\max_{\beta \in \mathbb{R}^{n \times k}} -\sum_{i=1}^{n} [\Omega(\beta_i) - \Omega(y_i)] - G^*(X^\top (Y - \beta)) \text{ s.t. } \beta_i \in \mathrm{dom}(\Omega)$$

  with primal-dual link $W^\star = \nabla G^*(X^\top (Y - \beta^\star))$. Note that $Y \in \{0, 1\}^{n \times k}$ contains the labels in one-hot representation.

Duality gap

- Let $P(W)$ and $D(\beta)$ be the primal and dual objectives, respectively.
- For all W and β we have

$$D(\beta) \leq P(W)$$

- At the optima, we have

$$D(\beta^\star) = P(W^\star)$$

- $P(W) - D(\beta) \geq 0$ is called the duality gap and can be used as a certificate of optimality.

Outline

Block coordinate ascent

- Key idea: on each iteration, pick a block of variables $\beta_i \in \mathbb{R}^k$ and update only that block.
- If the block has a size of 1, this is called coordinate ascent.
- **Exact update:**

$$\beta_i \leftarrow \underset{\beta_i \in \mathrm{dom}(\Omega)}{\mathrm{argmin}} \; \Omega(\beta_i) - \Omega(y_i) + G^*(X^\top(Y - \beta)) \qquad i \in \{1, \dots, n\}$$

- Possible schemes for picking i: random, cyclic, shuffled cyclic

Block coordinate ascent

- The sub-problem can be too complicated in some cases.
- **Approximate update** (using a quadratic approximation around the current iterate β_i^t)

$$\begin{aligned}\beta_i &\leftarrow \operatorname*{argmin}_{\beta_i \in \mathrm{dom}(\Omega)} \Omega(\beta_i) - \langle \beta_i, u_i \rangle + \frac{\sigma_i}{2} \|\beta_i\|_2^2 \\ &= \mathrm{prox}_{\frac{1}{\sigma_i}\Omega}(u_i / \sigma_i)\end{aligned}$$

where $\sigma_i = \frac{\|x_i\|_2^2}{\lambda}$ and $u_i = \underbrace{\nabla G^*(X^\top (Y - \beta))}_{w} x_i + \sigma_i \beta_i^t$.

- Exact if both Ω and G^* are quadratic
- Enjoys a linear rate of convergence w.r.t. the primal objective if Ω and G are strongly-convex.

Proximal operators

- Squared loss: $\Omega(\beta_i) = \frac{1}{2}\|\beta_i\|_2^2$

$$\mathrm{prox}_{\tau\Omega}(\eta) = \operatorname*{argmin}_{\beta \in \mathbb{R}^k} \frac{1}{2}\|\beta - \eta\|_2^2 + \frac{\tau}{2}\|\beta\|_2^2 = \frac{\eta}{\tau + 1}$$

- Perceptron loss: $\Omega(\beta_i) = I_{\triangle^k}(\beta_i)$

$$\mathrm{prox}_{\tau\Omega}(\eta) = \operatorname*{argmin}_{p \in \triangle^k} \|p - \eta\|_2^2$$

- Multiclass hinge loss: $\Omega(\beta_i) = I_{\triangle^k}(\beta_i) - \langle \beta_i, v_i \rangle$

$$\mathrm{prox}_{\tau\Omega}(\eta) = \operatorname*{argmin}_{p \in \triangle^k} \|p - (\eta + \tau v_i)\|_2^2$$

where $v_i = 1 - y_i$ and $y_i \in \{e_1, \ldots, e_k\}$ is the correct label.

Outline

Summary

- Conjugate functions are a powerful abstraction.
- Smoothing techniques are enabled by the duality between smoothness and strong convexity.
- The dual can often be easier to solve than the primal.
- If the dual is quadratic and the constraints are decomposable, dual block coordinate ascent is very well suited.

References

- Some figures and contents are borrowed from L. Vandenberghe's lectures "Subgradients" and "Conjugate functions".

- The block coordinate ascent algorithm used is described in

  Accelerated Proximal Stochastic Dual Coordinate Ascent for Regularized Loss Minimization
  Shai Shalev-Shwartz, Tong Zhang
  2013

- as well as in

  Learning with Fenchel-Young losses
  Mathieu Blondel, Andre Martins, Vlad Niculae
  2019

Lab work

Implement BDCA for the squared loss and the multiclass hinge loss.

- Primal objective

$$P(W) = \sum_{i=1}^{n} L_\Omega(\theta_i, y_i) + G(W) \qquad \theta_i = W^\top x_i \in \mathbb{R}^k, y_i \in \mathbb{R}^k$$

- Dual objective

$$D(\beta) = -\sum_{i=1}^{n} [\Omega(\beta_i) - \Omega(y_i)] - G^*(X^\top (Y - \beta)) \text{ s.t. } \beta_i \in \mathrm{dom}(\Omega)$$

  with primal-dual link $W^\star = \nabla G^*(X^\top (Y - \beta^\star))$. Note that $Y \in \{0, 1\}^{n \times k}$ contains the labels in one-hot representation.

- Duality gap $P(W) - D(\beta)$
- For G, use the squared L_2 norm

Lab work

- Approximate block update

$$\beta_i \leftarrow \mathrm{prox}_{\frac{1}{\sigma_i}\Omega}(u_i/\sigma_i)$$

where $\sigma_i = \frac{\|x_i\|_2^2}{\lambda}$ and $u_i = \underbrace{\nabla G^*(X^\top(Y-\beta))}_{w} x_i + \sigma_i \beta_i^t$.

- Use cyclic block selection
- See "Proximal operators" slide for prox expressions
- See "Fenchel-Young losses" slide for L_Ω and Ω expressions
- See "Examples of regularizer" slide for G^* and ∇G^* expressions