Matrix Algebra and Optimization for Statistics and Machine Learning

Yiyuan She

Department of Statistics, Florida State University

- Bregman divergence, mirror descent, and accelerations

Bregman divergence

- Let $f : \Omega \to \mathbb{R}$ be continuously differentiable and strictly convex. Then

$$\mathbf{D}_f(x, y) \triangleq f(x) - f(y) - \langle \nabla f(y), x - y \rangle, \forall x, y \in \Omega$$

- Why called a "divergence"? $\mathbf{D}_f(x, y) > 0$ unless $y = x$
- We can generalize the Bregman notation to $\boldsymbol{\Delta}_f$, for any f directionally differentiable in the direction $x - y$

Examples

- $\mathbf{D}_{(\cdot)^2/2}(x,y) = \|x-y\|_2^2/2 \equiv \mathbf{D}_2$ (metric)
- Let $\varphi(p) = \sum p_i \log p_i$ (negentropy). Then we get the un-normalized or normalized KL divergence

$$\begin{aligned}\mathbf{D}_\varphi(p,q) &= \sum p_i \log p_i - \sum q_i \log q_i - \langle 1 + \log q, p - q\rangle \\ &= \sum p_i \log(p_i/q_i) - p_i + q_i \\ &= \sum p_i \log(p_i/q_i) \text{ if } \sum p_i = \sum q_i = 1\end{aligned}$$

- $\varphi = -\sum \log p_i$ gives Itakura-Saito: $\mathbf{D}_\varphi(p,q) = \sum -\log p_i + \log q_i + (p_i - q_i)/q_i = \sum p_i/q_i - \log p_i/q_i - 1$

Properties

- $\mathbf{D}_\varphi(\cdot, y)$ is strictly convex given any $y \in \Omega$
- $\nabla_x \mathbf{D}_\varphi(x, y) = \nabla\varphi(x) - \nabla\varphi(y)$
- $\mathbf{D}_{a\varphi+\phi}(x, y) = a\mathbf{D}_\varphi(x, y) - \mathbf{D}_\phi(x, y)$
- $\mathbf{D}_{\mathbf{D}_\varphi(\cdot,z)}(x, y) = \mathbf{D}_\varphi(x, y)$ or the 3-point property:
  $\mathbf{D}_\varphi(x, y) = \mathbf{D}_\varphi(x, z) - \mathbf{D}_\varphi(y, z) - \langle x - y, \nabla\varphi(y) - \nabla\varphi(z)\rangle$
- Let x^o be a local minimizer of $f(x)$. Then
  $f(x) - f(x^o) \geq \mathbf{\Delta}_f(x, x^o)$ for any $x \in \mathrm{dom} f$

Strict convexity, conjugate & Bregman

- Recall $\varphi^*(y) = \sup_{x \in \Omega} \langle y, x \rangle - \varphi(x)$. For simplicity, let $\varphi \in \mathcal{C}^{(1)}$ be strictly convex and $\Omega := \mathrm{dom}\varphi = \mathbb{R}^n$.
- Given y, the problem has a unique solution x satisfying $y = \nabla\varphi(x)$ or $x = (\nabla\varphi)^{-1}(y)$ (well-defined!)
  - Notice the one-to-one mapping
- Naturally, for any x, let $x^* = \nabla\varphi(x)$ (dual point), then

$$\varphi^*(x^*) + \varphi(x) = \langle x, x^* \rangle$$

[No need of *strict* convexity. The result holds generally if f is closed and convex, and $x^* \in \partial\varphi(x)$.]

- $\nabla\varphi^*(y) = x^* = (\nabla\varphi)^{-1}(y)$, $\forall y$ or $\nabla\varphi(\nabla\varphi^*(\cdot)) = Id$. Similarly, from $\varphi^{**} = \varphi$, we know $\nabla\varphi^*(\nabla\varphi(\cdot)) = Id$
- Now it is easy to show that

$$\mathbf{D}_{\varphi}(p, q) = \mathbf{D}_{\varphi^*}(q^*, p^*)$$

where $p^* = \nabla\varphi(p)$, $q^* = \nabla\varphi(q)$

Bregman divergence & exponential family

- For every distribution in the (regular) exponential family, there exists an associated Bregman divergence
- Let $p(x|\theta) = \exp(x^T\theta - b(\theta))a(x)$ with b the cumulant
- Recall that $\mu(\theta) = \mathbb{E}[x] = b'(\theta)$. Define $\varphi = b^*$. Then

$$-\log p(x|\theta) = \mathbf{D}_{\varphi}(x, \mu(\theta)) + c(x)$$

where $c(x)$ does not depend on θ

- The proof is straightforward ($b \in \mathcal{C}^{(1)}$, strictly convex):

$$\begin{aligned}
-\langle x, \theta \rangle + b(\theta) &= -\langle x, \theta \rangle + \{\langle \mu, \theta \rangle - \varphi(\mu)\} \\
&= -\varphi(\mu) - \langle x - \mu, \theta \rangle \\
&= -\varphi(\mu) - \langle x - \mu, \nabla \varphi(\mu) \rangle \\
&= \mathbf{D}_\varphi(x, \mu) - \varphi(x)
\end{aligned}$$

- Another perspective: $\min_{\theta,\eta} -\langle x, \theta \rangle + b(\eta)$ s.t. $\theta = \eta \Rightarrow L(\theta, \eta, \mu) = -\langle x, \theta \rangle + b(\eta) + \langle \mu, \theta - \eta \rangle$. μ: dual variable.
- **Primal-dual**: $\min_\theta \max_\mu -\langle x - \mu, \theta \rangle - \varphi(\mu)$, since

$$\min_{\eta} L(\theta, \eta, \mu) = -\langle x - \mu, \theta \rangle - \varphi(\mu)$$

Examples

- Gaussian:

$$b(\theta) = \theta^2/2, \varphi(\mu) = \mu^2/2, \mathbf{D}_\varphi(x, \mu) = (x - \mu)^2/2$$

- Multinomial: $(x_1, \ldots, x_m) \sim m(1, p_1, \ldots, p_m)$

$$b(\theta) = \log \sum \exp(\theta_i), \varphi(\mu) = \sum (\mu_i \log \mu_i) \iota_{1^T \mu = 1}$$

$$\mathbf{D}_\varphi(x, \mu) = \mathbf{D}_{\mathrm{KL}}(x, \mu) = \sum x_i \log(x_i/\mu_i)$$

- Poisson:

$$b(\theta) = \exp(\theta), \varphi(\mu) = \mu \log \mu - \mu \Rightarrow$$
$$\mathbf{D}_\varphi(x, \mu) = \mathbf{D}_{\mathrm{KL}}(x, \mu) = x \log \frac{x}{\mu} - x + \mu$$

- Exponential:

$$p(x|\theta) = (-\theta) \exp(\theta x) 1_{x \geq 0}, b(\theta) = -\log(-\theta),$$
$$\varphi(\mu) = -\log \mu - 1, \mathbf{D}_\varphi(x, \mu) = \mathbf{D}_{\mathrm{IS}} = \frac{x}{\mu} - \log \frac{x}{\mu} - 1$$

Mirror descent

- Let f be convex. Recall GD: $\beta^{t+1} = \beta^t - \alpha_t \nabla f(\beta^t)$
- Given a strictly convex function φ, MD proceeds by

$$\begin{aligned}\beta^{t+1} &= \nabla\varphi^*(\nabla\varphi(\beta^t) - \alpha_t \nabla f(\beta^t)) \\ &= (\nabla\varphi)^{-1}(\nabla\varphi(\beta^t) - \alpha_t \nabla f(\beta^t))\end{aligned}$$

- Mirror: Map, run GD in the dual space, and map back
- When f is not smooth, choose a subgradient $\in \partial f(\beta^t)$

Surrogate and proximity

- Let $f = l + P$. Consider a surrogate by linearization

$$
\begin{aligned}
g(\beta, \beta^-) &= l(\beta) + (\rho \mathbf{D}_\varphi - \mathbf{\Delta}_l)(\beta, \beta^-) + P(\beta) \\
&= l(\beta^-) + \langle \nabla l(\beta^-), \beta - \beta^- \rangle + \rho \mathbf{D}_\varphi(\beta, \beta^-) + P(\beta)
\end{aligned}
$$

- When $\mathbf{D}_\varphi(\beta, \beta^-) = \mathbf{D}_2(\beta, \beta^-) = \|\beta - \beta^-\|_2^2/2$, $\beta^{t+1} = \arg\min_\beta g(\beta, \beta^t)$ gives proximal gradient descent
- Interestingly, β^{t+1} can be obtained stepwise

$$
\begin{aligned}
\gamma^{t+1} &= (\nabla \varphi)^{-1}(\nabla \varphi(\beta^t) - \nabla l(\beta^t)/\rho) \\
\beta^{t+1} &= \arg\min \mathbf{D}_\varphi(\beta, \gamma^{t+1}) + P(\beta)
\end{aligned}
$$

- This is because

$$\begin{aligned}
&\langle \alpha \nabla l(\beta^-), \beta \rangle + \mathbf{D}_\varphi(\beta, \beta^-) \\
&= \varphi(\beta) - \langle \nabla\varphi(\beta^-) - \alpha \nabla l(\beta^-), \beta \rangle - \varphi(\beta^-) + \langle \nabla\varphi(\beta^-), \beta^- \rangle \\
&= \varphi(\beta) - \langle \nabla\varphi(\gamma^-), \beta \rangle - \varphi(\beta^-) + \langle \nabla\varphi(\beta^-), \beta^- \rangle \\
&= \mathbf{D}_\varphi(\beta, \gamma^-) - \langle \nabla\varphi(\gamma^-), \gamma^- \rangle - \varphi(\beta^-) + \varphi(\gamma^-) + \langle \nabla\varphi(\beta^-), \beta^- \rangle
\end{aligned}$$

  where $\alpha = 1/\rho$, $\nabla\varphi(\gamma^-) = \nabla\varphi(\beta^-) - \alpha\nabla l(\beta^-)$
- $P(\beta) = \iota_C$: the second step gives a Bregman projection
- GD, projected GD, proximal GD are all special cases

Example: exponential gradient descent

- Consider a problem on the probability simplex: $\min l(\beta)$ s.t. $\beta \in C = \{\beta_j \geq 0 \, \forall j, 1^T\beta = 1\}$ (e.g., **EL**)
- We can surely use Lagrangian, but can we maintain the constraints automatically when doing the update?
- Choose $\varphi(\beta) = \sum \beta_j \log \beta_j - \beta_j$ and so $\mathbf{D}_\varphi = \mathbf{D}_{\text{KL}}$
- $\nabla\varphi = \log \beta$, $(\nabla\varphi)^{-1}(\cdot) = \exp(\cdot)$ (componentwise)
- $\log \gamma^{t+1} = \log \beta^t - \alpha_t \nabla l(\beta^t), \gamma^{t+1} = \beta^t \circ \exp(-\alpha_t \nabla l(\beta^t))$

- Therefore, if $\beta_j^0 \geq 0$ for any j, so are β^t, $t = 1, 2, \ldots$
- The multiplicative update is also widely seen in nonnegative matrix factorization (NMF)
- How about the Bregman projection $\min_{\beta \in C} \mathbf{D}_\varphi(\beta, \gamma)$?
- Lagrangian gives $\beta^o = \gamma / \sum \gamma_j$ (normalization)
- Therefore, the complete mirror descent algorithm is

$$\gamma^{t+1} = \beta^t \circ \exp(-\alpha_t \nabla l(\beta^t)), \quad \beta^{t+1} = \frac{\gamma^{t+1}}{\sum \gamma_j^{t+1}}$$

- [Analysis: use the 1-strong convexity of φ w.r.t. $\|\cdot\|_1$.]

Online learning

- Consider a game between a player against an adversary: At round t, (i) the player chooses $a_t \in \mathcal{A}$; (ii) the adversary picks a function l_t; (iii) the player suffers a loss $l_t(a_t)$; (iv) the player observes l_t
- Regression: learner – β_t, adversary – $l_t(\cdot) = l(\cdot; x_t, y_t)$
- Goal (for player): minimize the (cumulative) **regret**

$$R_T = \sum_{t=1}^{T} l_t(a_t) - \inf_{a \in \mathcal{A}} \sum_{t=1}^{T} l_k(a)$$

- The infimum may be taken among N fixed experts

- At the player's side, there is no knowledge of how to pick l_t (no model). Let's assume they are all convex.
- A strategy by online mirror descent: Choose φ.

$$a_1 = \arg\min_{a \in \mathcal{A}} \varphi(a),$$

$$\begin{cases} w_{t+1} = \nabla \varphi^*(\nabla \varphi(a_t) - \eta \nabla l_t(a_t)), \\ a_{t+1} = \arg\min_{a \in \mathcal{A}} \mathbf{D}_\varphi(a, w_{t+1}) \end{cases}$$

- Example: $\mathcal{A}$: simplex, exponentially weighting

- We can explain it using the linearized current loss

$$a_{t+1} = \arg\min_{a \in \mathcal{A}} l_t(a_t) + \langle \nabla l_t(a_t), a - a_t \rangle + \frac{1}{\eta} \mathbf{D}_\varphi(a, a_t)$$

- Follow the regularized leader
- Need to derive regret bounds to get a wise choice of η. [If $R_T \sim \sqrt{T}$, the average regret vanishes eventually.]

Nesterov's Accelerations

- Mirror descent type algorithms are often "effortless" in each iteration but converge slower than Newton
- Can we accelerate these first-order algorithms without incurring much additional cost (per iteration)?
- **Accelerated** gradients (Nesterov 83, 88, 05, Beck & Teboulle 08) can achieve the rate of $O(1/T^2)$
- [We can indeed adapt them to **nonconvex** settings with a careful control of relaxation and step size.]

Two scenarios

- Problem: $\min f(\beta)$. Assume convexity for simplicity.
- The 1st acceleration is to accelerate gradient descent

$$g(\beta, \gamma) = f(\beta) - \mathbf{\Delta}_{\psi_0}(\beta, \gamma) + \rho \mathbf{D}_2(\beta, \gamma)$$

  - GD: $g = f - \mathbf{\Delta}_f + \rho \mathbf{D}_2$, proximal: $f - \mathbf{\Delta}_l + \rho \mathbf{D}_2$
- The 2nd acceleration applies to mirror descent

$$g(\beta, \gamma) = f(\beta) - \mathbf{\Delta}_{\psi_0}(\beta, \gamma) + \rho \mathbf{\Delta}_{\phi}(\beta, \gamma)$$

where ϕ is strongly convex

The first acceleration scheme

- For problems in Scenario 1, consider

$$\gamma^{(t)} = \beta^{(t)} + \theta_t(\theta_{t-1}^{-1} - 1)(\beta^{(t)} - \beta^{(t-1)}),$$
$$\beta^{(t+1)} = \arg\min_{\beta}\{f(\beta) - \Delta_{\psi_0}(\beta, \gamma^{(t)}) + \rho_t \mathbf{D}_2(\beta, \gamma^{(t)})\},$$

- Momentum-based update using an auxiliary sequence
- The key lies in picking $\{\theta_t\}$, $\{\rho_t\}$ properly:

$$(\rho_t \mathbf{D}_2 - \Delta_{\psi_0})(\beta^{(t+1)}, \gamma^{(t)}) + (1-\theta_t)\Delta_{\psi_0}(\beta^{(t)}, \gamma^{(t)}) \geq 0$$
$$\frac{\theta_t^2}{1-\theta_t} = \frac{\rho_{t-1}\theta_{t-1}^2}{\rho_t},\ \ \theta_t \geq 0,\ \ \rho_t > 0,\ \ t \geq 1, \theta_0 = 1$$

- GD: Let $\psi_0 = f$ which is convex and has L-Lipschitz continuous gradient. Then the following choices suffice

$$\rho_t = L, \quad \theta_{t+1} = (\sqrt{\theta_t^4 + 4\theta_t^2} - \theta_t^2)/2$$

- We can prove for any β

$$f(\beta^{T+1}) - f(\beta) + \min_{0 \le t \le T} \mathbf{\Delta}_{\psi_0}(\beta, \gamma^t)$$
$$\le \{\theta_T^2 \rho_T \vee \frac{1}{\sum_0^T 1/(\theta_t \rho_t)}\} \mathbf{D}_2(\beta, \beta^0) = \mathcal{O}(\frac{L}{T^2}) \mathbf{D}_2(\beta, \beta^0)$$

The second acceleration

- This time we use two auxiliary sequences

$$\gamma^{(t)} = (1-\theta_t)\beta^{(t)} + \theta_t \alpha^{(t)},$$
$$\alpha^{(t+1)} = \arg\min_{\beta} f(\beta) - \Delta_{\psi_0}(\beta, \gamma^{(t)}) + \theta_t \rho_t \boldsymbol{\Delta}_{\phi}(\beta, \alpha^{(t)})$$
$$\beta^{(t+1)} = (1-\theta_t)\beta^{(t)} + \theta_t \alpha^{(t+1)}$$

- Stepsize & relaxation: $\theta_t^2 \rho_t \boldsymbol{\Delta}_{\phi}(\alpha^{(t+1)}, \alpha^{(t)}) - \Delta_{\psi_0}(\beta^{(t+1)}, \gamma^{(t)}) + (1-\theta_t)\Delta_{\psi_0}(\beta^{(t)}, \gamma^{(t)}) \geq 0$; θ_t still as before

- Assume $\nabla\psi_0$ is Lipschitz: $\Delta_{\psi_0} \leq L_{\psi_0}\mathbf{D}_2$ and ϕ is strongly convex: $\mathbf{D}_\phi \geq \sigma\mathbf{D}_2$.
- Then as long as $\rho_t \geq L_{\psi_0}/\sigma$, the condition is satisfied
- Similarly, we can show for any β

$$\frac{f(\beta^{(T+1)}) - f(\beta)}{\theta_T^2\rho_T} + T \operatorname*{avg}_{0\leq t\leq T}\left\{\frac{\Delta_{\psi_0}(\beta,\gamma^{(t)})}{\theta_t\rho_t}\right\} \leq \boldsymbol{\Delta}_\phi(\beta,\alpha^{(0)})$$

So $f(\beta^{(T+1)}) - f(\beta) + \min_{t\leq T}\Delta_{\psi_0}(\beta,\gamma^{(t)}) \leq \mathcal{O}(\frac{L_{\psi_0}}{\sigma T^2})$.