DSCC/CSC 435 & ECE 412 Optimization for Machine Learning	Lecture 12

Optimization in Relative Scale

Lecturer: Jiaming Liang *November 2, 2023*

1 Relative smoothness and relative strong convexity

There are many differentiable convex functions in practice that do not satisfy a uniform smoothness condition, e.g., the D-optimal design problem. Given a matrix $H \in \mathbb{R}^{m\times n}$ of rank m, where $n \geq m+1$, the D-optimal design problem is

$$\min_{x\in\Delta_n} \left\{ f(x) := -\ln\det\left(HXH^\top\right)\right\}$$

where $X = \mathrm{Diag}(x)$. In statistics, the D-optimal design problem corresponds to maximizing the determinant of the Fisher information matrix $\mathbb{E}[HH^\top]$. In computational geometry, D-optimal design arises as a Lagrangian dual problem of the minimum volume covering ellipsoid problem.

We are interested in solving a constrained problem

$$\min_{x\in Q} f(x),$$

where f is closed and convex and Q is a closed and convex set. We do not assume that f is uniformly smooth or strongly convex, but instead we resort to the following notions of relative smoothness and strong convexity.

Definition 1. *We say f is L-smooth relative to h on Q if for any $x, y \in \operatorname{int} Q$, there is a scalar L for which*

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

where D_h is the Bregman divergence of h.

Definition 2. *We say f is μ-strongly convex relative to h on Q if for any $x, y \in \operatorname{int} Q$, there is a scalar $\mu \geq 0$ for which*

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

Note that h does not need to strongly nor strictly convex. We refer to h as the reference function.

In the case when both f and h are twice differentiable, f is both μ-strongly convex and L-smooth relative to h can be written as

$$\mu\nabla^2 h(x) \preceq \nabla^2 f(x) \preceq L\nabla^2 h(x) \text{ for all } x\in \operatorname{int} Q.$$

Lemma 1. *The following conditions are equivalent:*

(a) $f(\cdot)$ *is* L*-smooth relative to* $h(\cdot)$*;*

(b) $Lh(\cdot) - f(\cdot)$ *is a convex function on* Q*;*

(c) under twice differentiability $\nabla^2 f(x) \preceq L\nabla^2 h(x)$ *for any* $x \in \operatorname{int} Q$*;*

(d) $\langle \nabla f(x) - \nabla f(y), x - y\rangle \leq L\langle \nabla h(x) - \nabla h(y), x - y\rangle$ *for all* $x, y \in \operatorname{int} Q$*.*

Lemma 2. *The following conditions are equivalent:*

(a) $f(\cdot)$ *is* μ*-strongly convex relative to* $h(\cdot)$*;*

(b) $f(\cdot) - \mu h(\cdot)$ *is a convex function on* Q*;*

(c) under twice differentiability $\nabla^2 f(x) \succeq \mu\nabla^2 h(x)$ *for any* $x \in$ *int* Q*;*

(d) $\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \mu\langle \nabla h(x) - \nabla h(y), x - y\rangle$ *for all* $x, y \in \operatorname{int} Q$*.*

Example 1. Suppose that f is a twice-differentiable convex function on $Q := \mathbb{R}^n$ and let $\left\|\nabla^2 f(x)\right\|$ denote the operator norm of $\nabla^2 f(x)$ with respect to the ℓ_2-norm on $\mathbb{R}^n$. Suppose that $\left\|\nabla^2 f(x)\right\| \leq p_r\left(\|x\|_2\right)$ where $p_r(\alpha) = \sum_{i=0}^r a_i\alpha^i$ is an r-degree polynomial of α. Let

$$h(x) := \frac{1}{r+2}\|x\|_2^{r+2} + \frac{1}{2}\|x\|_2^2.$$

Then the following lemma shows that f is L-smooth relative to h with a certain L.

Lemma 3. *Let* L *be such that* $p_r(\alpha) \leq L\left(1 + \alpha^r\right)$ *for* $\alpha \geq 0$*. Then* f *is* L*-smooth relative to* h*.*

Proof. Calculation gives the gradient of h

$$\nabla h(x) = \|x\|^{r+1}\frac{x}{\|x\|} + x = \|x\|^r x + x,$$

and its Hessian

$$\begin{aligned}\nabla^2 h(x) &= \|x\|^r I + xr\|x\|^{r-1}\frac{x^\top}{\|x\|} + I = (1 + \|x\|^r)I + r\|x\|^{r-2}xx^\top \\ &\succeq (1 + \|x\|^r)I \succeq \frac{1}{L}p_r(\|x\|)I \succeq \frac{1}{L}\nabla^2 f(x),\end{aligned}$$

where the last two relations follow from the assumptions on f and p_r. So, f is L-smooth relative to h by Lemma 1(c). $\square$

Example 2. D-optimal design In this case, we choose h to be the logarithmic barrier function, namely,

$$h(x) := -\sum_{j=1}^{n} \ln(x_j)$$

defined on the positive orthant $\mathbb{R}_{++}$.

Lemma 4. *The f in D-optimal design is 1-smooth relative to h on $\mathbb{R}_{++}$.*

Proof. The gradient and Hessian of h are

$$\nabla f(x) = \operatorname{diag}(-C), \quad C = H^\top (HXH^\top)^{-1} H,$$

and

$$\nabla^2 f(x) = C \circ C$$

where $\circ$ denotes the Hadamard product. Let $U = HX^{1/2}$, then

$$U^\top (UU^\top)^{-1} U \preceq I$$

since the left side of this matrix inequality is a projection operator. Then, we have

$$X^{\frac{1}{2}} H^\top \left(HXH^\top\right)^{-1} HX^{\frac{1}{2}} \preceq I.$$

Multiplying this matrix inequality on the left and right by $X^{-\frac{1}{2}}$, then we have

$$C \preceq X^{-1}.$$

Moreover, we get

$$\nabla^2 f(x) = C \circ C \preceq C \circ X^{-1} \preceq X^{-1} \circ X^{-1} = X^{-2} = \nabla^2 h(x)$$

where the first and the second matrix inequalities above follows from the fact that $C \preceq X^{-1}$ and the Hadamard product of two symmetric positive semidefinite matrices is also a symmetric positive semidefinite matrix. The result then follows using Lemma 1(c). $\square$

2 Algorithms

2.1 Primal gradient method

Algorithm 1 Primal gradient method with reference h

Input: Initial point $x_0 \in Q$, L and h satisfying Definition 1 be given
for $k \geq 0$ **do**
  Compute $x_{k+1} = \operatorname{argmin}_{x \in Q} \{\ell_f(x; x_k) + LD_h(x, x_k)\}$.
end for

The following lemma is a stronger version of the three-points lemma.

Lemma 5. *Suppose ϕ is convex and let*

$$z^{+} := \arg\min_{x\in Q} \left\{\phi(x) + D_h(x,z)\right\},$$

then for all $x \in Q$

$$\phi(x) + D_h(x,z) \geq \phi\left(z^{+}\right) + D_h\left(z^{+}, z\right) + D_h\left(x, z^{+}\right).$$

Theorem 1. *If f is L-smooth and μ-strongly convex ralative to h, then $\{x_k\}$ generated by the primal gradient method satisfies*

$$f\left(x_k\right) - f(x_*) \leq \frac{\mu D_h\left(x_*, x_0\right)}{\left(1+\frac{\mu}{L-\mu}\right)^k - 1} \leq \frac{L-\mu}{k} D_h\left(x_*, x_0\right).$$

Proof. It follows from Definitions 1 and 2 that

$$\begin{aligned} f\left(x_k\right) &\leq f\left(x_{k-1}\right) + \left\langle \nabla f\left(x_{k-1}\right), x_k - x_{k-1}\right\rangle + L D_h\left(x_k, x_{k-1}\right) \\ &\leq f\left(x_{k-1}\right) + \left\langle \nabla f\left(x_{k-1}\right), x - x_{k-1}\right\rangle + L D_h\left(x, x_{k-1}\right) - L D_h\left(x, x_k\right) \\ &\leq f(x) + (L-\mu) D_h\left(x, x_{k-1}\right) - L D_h\left(x, x_k\right). \end{aligned}$$

Taking $x = x_{k-1}$ in the above inequality, we know $\{f(x_k)\}$ is monotone. Multiplying $\left(\frac{L}{L-\mu}\right)^i$ to the above inequality and summing, we have

$$\sum_{i=1}^{k} \left(\frac{L}{L-\mu}\right)^i f\left(x_i\right) \leq \sum_{i=1}^{k} \left(\frac{L}{L-\mu}\right)^i f(x) + L D_h\left(x, x_0\right) - \left(\frac{L}{L-\mu}\right)^k L D_h\left(x, x_k\right),$$

and thus

$$\left(\sum_{i=1}^{k} \left(\frac{L}{L-\mu}\right)^i\right)\left(f\left(x_k\right) - f(x)\right) \leq L D_h\left(x, x_0\right) - \left(\frac{L}{L-\mu}\right)^k L D_h\left(x, x_k\right) \leq L D_h\left(x, x_0\right).$$

It follows from the monotonicity of $\{f(x_k)\}$ that

$$f\left(x_k\right) - f(x) \leq \frac{\mu D_h\left(x, x_0\right)}{\left(1+\frac{\mu}{L-\mu}\right)^k - 1}.$$

The first inequality of the theorem follows from the above inequality with $x = x_*$ and the second inequality holds by simple algebra. □

Example 1 continued. A key step in Algorithm 1 is to solve the subproblem with D_h. We know specify how to solve it in Example 1. The subproblem can be abstracted as

$$\min_{x\in\mathbb{R}^n} \langle c, x\rangle + \frac{1}{r+2}\|x\|_2^{r+2} + \frac{1}{2}\|x\|^2.$$

Its first-order optimality condition is

$$c + (1 + \|x\|_2^r)\, x = 0.$$

Clearly, we know $x = -\theta c$ for some $\theta \geq 0$, and it remains to simply determine the value of the nonnegative scalar θ. If $c = 0$, then $x = 0$ satisfies the optimality conditions. For $c \neq 0$, notice from above that θ must satisfy

$$1 - \theta - \|c\|_2^r \cdot \theta^{r+1} = 0,$$

which is a univariate polynomial in θ with a unique positive root.

Example 2 continued. The subproblem in D-optimal design is

$$\min_{x \in \Delta_n} \langle c, x \rangle - \sum_{j=1}^{n} \ln(x_j),$$

and its optimality condition is

$$x \geq 0, \quad \sum_{j=1}^{n} x_j = 1, \quad c - X^{-1}\mathbf{1} = -\theta \mathbf{1},$$

where θ is the lagrange multiplier. We thus have

$$x_j = \frac{1}{c_j + \theta}$$

and can solve

$$\sum_{j=1}^{n} \frac{1}{c_j + \theta} - 1 = 0$$

for θ.

2.2 Dual averaging method

Algorithm 2 Dual averaging method with reference h

Input: Initial point $x_0 = \operatorname{argmin}_{x \in Q} h(x)$, L, μ and h satisfying Definitions 1 and 2 be given

for $k \geq 0$ **do**

Compute $a_{k+1} = \frac{1}{L-\mu}\left(\frac{L}{L-\mu}\right)^k$ and

$$x_{k+1} = \operatorname{argmin}_{x \in Q}\{h(x) + \sum_{i=0}^{k} a_{i+1}\left(f\left(x^i\right) + \left\langle \nabla f\left(x^i\right), x - x_i \right\rangle + \mu D_h\left(x, x_i\right)\right)\}.$$

end for

Theorem 2. *If f is L-smooth and μ-strongly convex ralative to h, then $\{x_k\}$ generated by the dual averaging method satisfies*

$$\min_{1\le i\le k} f(x_i)-f(x_*)\le \frac{\mu[h(x_*)-h(x_0)]}{\left(1+\frac{\mu}{L-\mu}\right)^k-1}\le \frac{L-\mu}{k}[h(x_*)-h(x_0)].$$

Proof. We first define $\psi_0(x)=h(x)$ and for $k\ge 1$,

$$\psi_k(x):=h(x)+\sum_{i=0}^{k-1} a_{i+1}\left(f(x_i)+\langle\nabla f(x_i), x-x_i\rangle+\mu D_h(x,x_i)\right)$$

and $\psi_k^*:=\min_{x\in Q}\psi_k(x)$. Thus, we have $x_k=\arg\min_{x\in Q}\psi_k(x)$ and $\psi_k(x_k)=\psi_k^*$. It follows from the above definition and the relative strong convexity that

$$\psi_k^*\le h(x)+A_k f(x) \tag{1}$$

where

$$A_k:=\sum_{i=0}^{k-1} a_{i+1}=\frac{1}{\mu}\left[\left(1+\frac{\mu}{L-\mu}\right)^k-1\right].$$

Observe that the function ψ_k is a sum of a linear function and the reference function h multiplied by the coefficient $1+\mu A_k$. Therefore $(1+\mu A_k)\,h$ and ψ_k define the same Bregman distance, i.e., for any $x\in Q$ it holds that

$$(1+\mu A_k)\,D_h(x,x_k)=D_{\psi_k}(x,x_k)=\psi_k(x)-\psi_k(x_k)-\langle\nabla\psi_k(x_k), x-x_k\rangle\le\psi_k(x)-\psi_k^*.$$

The above inequality with $x=x_{k+1}$ and the definition of ψ_{k+1} imply that

$$\begin{aligned}
&\psi_{k+1}^*\\
&=\psi_{k+1}(x_{k+1})\\
&=\psi_k(x_{k+1})+a_{k+1}\left(f(x_k)+\langle\nabla f(x_k), x_{k+1}-x_k\rangle+\mu D_h(x_{k+1},x_k)\right)\\
&\ge\psi_k^*+a_{k+1}\left(f(x_k)+\langle\nabla f(x_k), x_{k+1}-x_k\rangle+\left(\mu+\frac{1}{a_{k+1}}(1+\mu A_k)\right)D_h(x_{k+1},x_k)\right).
\end{aligned}$$

Using the fact that

$$\mu+\frac{1}{a_{k+1}}(1+\mu A_k)=\frac{1+\mu A_{k+1}}{a_{k+1}}=\frac{1}{a_{k+1}}\left(\frac{L}{L-\mu}\right)^{k+1}=L$$

and the relative smoothness of f, we obtain for $k\ge 0$,

$$\psi_{k+1}^*\ge\psi_k^*+a_{k+1}f(x_{k+1}).$$

Summing up and using (1), we obtain

$$\sum_{i=0}^{k-1} a_{i+1}f(x_{i+1})\le\psi_k^*-h(x_0)\le h(x)+A_k f(x)-h(x_0).$$

The conclusions of the theorem immediately follow. $\square$