Quasi-Newton Methods

- Relationship with Conjugate Gradient Methods

We start by considering the Hestenes-Stiefel form of the nonlinear conjugate gradient method. Recalling that $\Delta\mathbf{x}_k = \alpha_k \mathbf{p}_k$, the search direction for this method is given by

$$\mathbf{p}_{k+1} = -\mathbf{g}_{k+1} + \frac{\mathbf{g}_{k+1}^T \Delta\mathbf{g}_k}{\Delta\mathbf{g}_k^T \mathbf{p}_k}\mathbf{p}_k = -\left(\mathbf{I} - \frac{\Delta\mathbf{x}_k \Delta\mathbf{g}_k^T}{\Delta\mathbf{g}_k^T \Delta\mathbf{x}_k}\right)\mathbf{g}_{k+1} \equiv -\hat{\mathbf{H}}_{k+1}\mathbf{g}_{k+1}. \quad (7)$$

This formula resembles a quasi-Newton iteration, but the matrix $\hat{\mathbf{H}}_{k+1}$ is neither symmetric nor positive definite. We could symmetrize it as $\hat{\mathbf{H}}_{k+1}^T \hat{\mathbf{H}}_{k+1}$, but this matrix does not satisfy the secant equation $\hat{\mathbf{H}}_{k+1}\Delta\mathbf{g}_k = \Delta\mathbf{x}_k$ and is, in any case, singular. An iteration matrix that is symmetric, positive definite, and satisfies the secant equation is given by

$$\mathbf{H}_{k+1} = \left(\mathbf{I} - \frac{\Delta\mathbf{x}_k \Delta\mathbf{g}_k^T}{\Delta\mathbf{g}_k^T \Delta\mathbf{x}_k}\right)\left(\mathbf{I} - \frac{\Delta\mathbf{g}_k \Delta\mathbf{x}_k^T}{\Delta\mathbf{g}_k^T \Delta\mathbf{x}_k}\right) + \frac{\Delta\mathbf{x}_k \Delta\mathbf{x}_k^T}{\Delta\mathbf{g}_k^T \Delta\mathbf{x}_k}. \quad (8)$$

Quasi-Newton Methods

- Relationship with Conjugate Gradient Methods

This matrix is exactly the one obtained by applying a single BFGS update (1) to the identity matrix. Hence, an algorithm whose search direction is given by $\mathbf{p}_{k+1} = -\mathbf{H}_{k+1}\mathbf{g}_{k+1}$, with $\mathbf{H}_{k+1}$ defined by (8), can be thought of as a "memoryless" BFGS method, in which the previous Hessian approximation is always reset to the identity matrix before updating it and where only the most recent correction pair $(\Delta\mathbf{x}_k, \Delta\mathbf{g}_k)$ is kept at every iteration. Alternatively, we can view the method as a variant of Algorithm 2 in which $m = 1$ and $\mathbf{H}_k^0 = \mathbf{I}$ at each iteration.

Quasi-Newton Methods

- Relationship with Conjugate Gradient Methods

A more direct connection with conjugate gradient methods can be seen if we consider the memoryless BFGS formula (8) in conjunction with an exact line search, for which $\mathbf{g}_{k+1}^T \mathbf{p}_k = 0$ for all k. We then obtain

$$\mathbf{p}_{k+1} = -\mathbf{H}_{k+1}\mathbf{g}_{k+1} = -\mathbf{g}_{k+1} + \frac{\mathbf{g}_{k+1}^T \Delta\mathbf{g}_k}{\Delta\mathbf{g}_k^T \mathbf{p}_k}\mathbf{p}_k, \tag{7}$$

which is none other than the Hestenes-Stiefel conjugate gradient method. Moreover, it is easy to verify that when $\mathbf{g}_{k+1}^T \mathbf{p}_k = 0$, the Hestenes-Stiefel formula reduces to the Polak-Ribiere formula. Even though the assumption of exact line searches is unrealistic, it is intriguing that the BFGS formula is related in this way to the Polak-Ribiere and Hestenes-Stiefel methods.

Majorization Minimization

- Basic framework

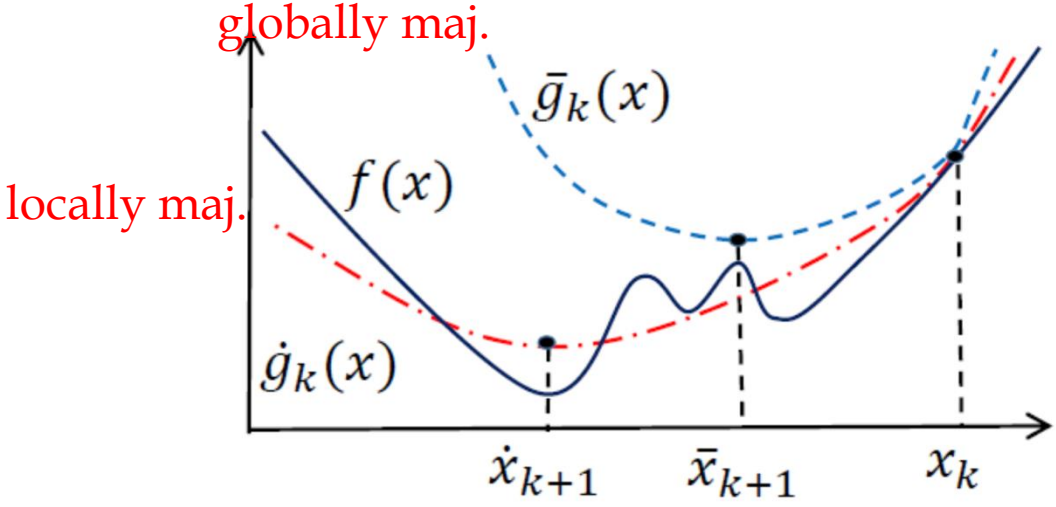
$$\min_{\mathbf{x}} \; f(\mathbf{x}), \quad s.t.\ \mathbf{x} \in \mathcal{C}.$$

$$\min_{\mathbf{x}} \; g_k(\mathbf{x}), \quad s.t.\ \mathbf{x} \in \mathcal{C}.$$

1. $f(\mathbf{x}) \leq g_k(\mathbf{x}), \forall \mathbf{x} \in \mathcal{C};$ globally majorant

2. $f(\mathbf{x}_k) = g_k(\mathbf{x}_k).$

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x}} g_k(\mathbf{x}) \Longrightarrow f(\mathbf{x}_{k+1}) \leq g_k(\mathbf{x}_{k+1}) \leq g_k(\mathbf{x}_k) = f(\mathbf{x}_k).$$



Majorization Minimization

- Basic convergence result

Definition 1 (First-order surrogate functions)**.** *A function* $g : \mathbb{R}^p \to \mathbb{R}$ *is a first-order surrogate function of* f *near* $\mathbf{k}_k$ *in* $\mathcal{C}$ *when*

(i) $g(\mathbf{x}') \geq f(\mathbf{x}')$ *for all minimizers* $\mathbf{x}'$ *of* g *over* $\mathcal{C}$*. When the more general condition* $g \geq f$ *holds, we say that* g *is a* majorizing surrogate*;*

(ii) the approximation error $h \triangleq g - f$ *is* L*-smooth,* $h(\mathbf{x}_k) = 0$*, and* $\nabla h(\mathbf{x}_k) = \mathbf{0}$*. We denote by* $\mathcal{S}_L(f, \mathbf{x}_k)$ *the set of first-order surrogate functions and by* $\mathcal{S}_{L,\rho}(f, \mathbf{x}_k) \subset \mathcal{S}_L(f, \mathbf{x}_k)$ *the subset of* ρ*-strongly convex surrogates.*

First-order surrogates are interesting because their approximation error – the difference between the surrogate and the objective – can be easily controlled.

Majorization Minimization

- Basic convergence result

Lemma 1 (Basic properties of first-order surrogate functions)**.** *Let g be a surrogate function in $\mathcal{S}_L(f, \mathbf{x}_k)$ for some $\mathbf{x}_k$ in $\mathcal{C}$. Define the approximation error $h \triangleq g - f$, and let $\mathbf{x}'$ be a minimizer of g over $\mathcal{C}$. Then, for all $\mathbf{x} \in \mathcal{C}$,*

(1) $|h(\mathbf{x})| \leq \frac{L}{2}\|\mathbf{x} - \mathbf{x}_k\|_2^2$;

(2) $f(\mathbf{x}') \leq f(\mathbf{x}) + \frac{L}{2}\|\mathbf{x} - \mathbf{x}_k\|_2^2$.

(3) Assume that g is ρ-strongly convex, i,e., g is in $\mathcal{S}_{L,\rho}(f, \mathbf{x}_k)$. Then $\forall \mathbf{x} \in \mathcal{C}$,

$$f(\mathbf{x}') + \frac{\rho}{2}\|\mathbf{x}' - \mathbf{x}\|_2^2 \leq f(\mathbf{x}) + \frac{L}{2}\|\mathbf{x} - \mathbf{x}_k\|_2^2.$$

Proof. (2) $\forall \mathbf{x} \in \mathcal{C}$, we have $f(\mathbf{x}') \leq g(\mathbf{x}') \leq g(\mathbf{x}) = f(\mathbf{x}) + h(\mathbf{x})$, ...
(3) When g is ρ-strongly convex, we use the following classical lower bound:

$$g(\mathbf{x}') + \frac{\rho}{2}\|\mathbf{x} - \mathbf{x}'\|_2^2 \leq g(\mathbf{x}).$$

Since $f(\mathbf{x}') \leq g(\mathbf{x}')$ and $g(\mathbf{x}) = f(\mathbf{x}) + h(\mathbf{x})$, we obain (3) from (1). □

Majorization Minimization

- Non-convex convergence analysis

For general non-convex problems, proving convergence to a global (or local) minimum is impossible in general, and classical analysis studies instead asymptotic stationary point conditions. To do so, we make the following mild assumption when f is non-convex:

(A) f is bounded below and for all $\mathbf{x}, \mathbf{x}' \in \mathcal{C}$, the directional derivative $\nabla f(\mathbf{x}; \mathbf{x}' - \mathbf{x})$ of f at $\mathbf{x}$ in the direction $\mathbf{x}' - \mathbf{x}$ exists.

Definition 3 (Stationary point)**.** *Consider a function $f : \mathcal{C} \subseteq \mathbb{R}^p \to \mathbb{R}$, where $\mathcal{C}$ is a convex set, such that f admits a directional derivative $\nabla f(\mathbf{x}; \mathbf{x}' - \mathbf{x})$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{C}$. We say that $\mathbf{x} \in \mathcal{C}$ is a stationary point if for all $\mathbf{x}' \in \mathcal{C}$, $\nabla f(\mathbf{x}; \mathbf{x}' - \mathbf{x}) \geq 0$.*

Majorization Minimization

- Non-convex convergence analysis

Definition 3 (Asymptotic stationary point)**.** *Under assumption (A), a sequence* $\{\mathbf{x}_n\}_{n\geq 0}$ *satisfies the asymptotic stationary point condition if*

$$\liminf_{n\to+\infty} \inf_{\mathbf{x}\in\mathcal{C}} \frac{\nabla f(\mathbf{x}_n; \mathbf{x}-\mathbf{x}_n)}{\|\mathbf{x}-\mathbf{x}_n\|_2} \geq 0. \tag{1}$$

Note that if f is differentiable on $\mathbb{R}^p$ and $\mathcal{C} = \mathbb{R}^p$, $\nabla f(\mathbf{x}_n; \mathbf{x}-\mathbf{x}_n) = \nabla f(\mathbf{x}_n)^T(\mathbf{x}-\mathbf{x}_n)$, and the condition (1) implies that the sequence $\{\nabla f(\mathbf{x}_n)\}_{n\geq 0}$ converges to $\mathbf{0}$. As noted, we recover the classical definition of critical points for the smooth unconstrained case.

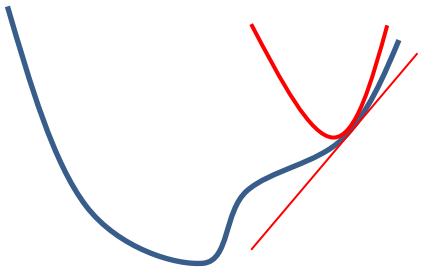
Majorization Minimization

- Non-convex convergence analysis

Proposition 3 (Non-convex analysis for MM)**.** *Assume that (A) holds and that the surrogates g_n are in $\mathcal{S}_L(f, \mathbf{x}_{n-1})$ and are either majorizing f or strongly convex. Then, $\{f(\mathbf{x}_n)\}_{n\geq 0}$ monotonically decreases, and $\{\mathbf{x}_n\}_{n\geq 0}$ satisfies the asymptotic stationary point condition.*

Majorization Minimization

- How to choose the majorant function?



Lipschitz Gradient Surrogate:

$$\min_{\mathbf{x}} \; g_k(\mathbf{x}), \quad s.t.\ \mathbf{x} \in \mathcal{C}.$$

$$g_k(\mathbf{x}) = f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + \frac{1}{2\alpha}\|\mathbf{x} - \mathbf{x}_k\|^2.$$

$$\mathbf{x}_{k+1} = \mathcal{P}_{\mathcal{C}}(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)).$$

projected gradient descent

$$(\mathcal{C} = \mathbb{R}^n)$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k).$$

gradient descent

Majorization Minimization

- How to choose the majorant function?
  - How to choose α?

$$f(\mathbf{x}_{k+1}) = f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)) < f(\mathbf{x}_k).$$ locally majorant

$$\Downarrow$$

If not satisfied $\alpha \leftarrow \mu\alpha$. $(\mu \in (0,1))$ backtracking

If f is L-smooth, i.e., $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$, then we may choose

$$\alpha = L^{-1}.$$

globally majorant

$$f(\mathbf{x}) \leq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + \frac{1}{2\alpha}\|\mathbf{x} - \mathbf{x}_k\|^2 \triangleq g_k(\mathbf{x}).$$

Majorization Minimization

- How to choose the majorant function?

Quadratic Surrogate:

$$g_k(\mathbf{x}) = f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + \frac{1}{2}(\mathbf{x} - \mathbf{x}_k)^T \mathbf{H}_k (\mathbf{x} - \mathbf{x}_k), \text{ where } \mathbf{H}_k \succ \nabla^2 f.$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{H}_k \nabla f. \quad (\mathcal{C} = \mathbb{R}^n)$$

Newton's method

Majorization Minimization

- How to choose the majorant function?

$$\min_{\mathbf{x}} \; g_k(\mathbf{x}), \quad s.t.\ \mathbf{x} \in \mathcal{C}.$$

$$g_k(\mathbf{x}) = f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + \frac{1}{2\alpha}\|\mathbf{x} - \mathbf{x}_k\|^2.$$

asymptotic smoothness

$g_k(\mathbf{x}) - f(\mathbf{x})$ is smooth.

$g_k(\mathbf{x}) \geq f(\mathbf{x}), \ \forall \mathbf{x}$

globally majorant

$$\lim_{k\to\infty} \nabla g_k(\mathbf{x}_k, \mathbf{d}) - \nabla f(\mathbf{x}_k; \mathbf{d}) = 0.$$

$$g_k(\mathbf{x}_{k+1}) \geq f(\mathbf{x}_{k+1}).$$

locally majorant

Relaxed Majorization Minimization

Robust Matrix Factorization: $\min_{\mathbf{U}\in\mathcal{C}_U, \mathbf{V}\in\mathcal{C}_V} \|\mathbf{W}\odot(\mathbf{M}-\mathbf{U}\mathbf{V}^T)\|_1 + R_u(\mathbf{U}) + R_v(\mathbf{V}).$

Majorization Minimization

- How to choose the majorant function?

$$\min_{\mathbf{x}} \; f(\mathbf{x}) \triangleq \tilde{f}(\boldsymbol{\theta}^T \mathbf{x}), \quad s.t.\ \mathbf{x} \in \mathcal{C}. \qquad \boxed{\tilde{f}(x) \text{ is convex}}$$

Jensen Surrogate:

$$g_k(\mathbf{x}) = \sum_{i=1}^{n} w_i \tilde{f}\left(\frac{\theta_i}{w_i}(x_i - x_{k,i}) + \boldsymbol{\theta}^T \mathbf{x}_k \right),$$

where $\mathbf{w} \in \mathbb{R}_+^n$, $\|\mathbf{w}\|_1$ and $w_i \neq 0$ whenever $\theta_i \neq 0$.

Majorization Minimization

- How to choose the majorant function?

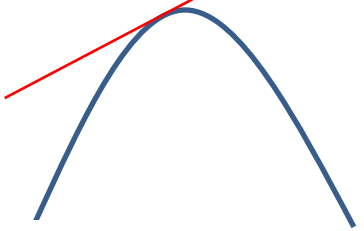
$$\min_{\mathbf{x}} \; f(\mathbf{x}) + h(\mathbf{x}), \quad s.t. \; \mathbf{x} \in \mathcal{C},$$

where f is convex and h is concave.

$$g_k(\mathbf{x}) = f(\mathbf{x}) + \langle \partial h(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + h(\mathbf{x}_k),$$

convex concave procedure (CCCP)

where ∂h is a super-gradient of h.



low-rankness regularizer

$$\min_{\mathbf{X}} \sum_{i=1}^{\min(m,n)} h(\sigma_i(\mathbf{X})) + f(\mathbf{X}), \text{ where } h \text{ is concave on } \mathbb{R}_+.$$

$$h(\sigma_i) \leq h(\sigma_i^k) + w_i^k(\sigma_i - \sigma_i^k), \quad w_i^k \in \partial h(\sigma_i).$$

Majorization Minimization

- How to choose the majorant function?

$$\min_{\mathbf{x}} \; f(\mathbf{x}), \quad s.t.\ \mathbf{x} \in \mathcal{C}, \quad \text{where } f(\mathbf{x}) = \min_{\mathbf{y} \in \mathcal{Y}} h(\mathbf{x}, \mathbf{y}).$$

Variational surrogate: $g_k(\mathbf{x}) = h(\mathbf{x}, \mathbf{y}_k^*), \quad \text{where } \mathbf{y}_k^* = \operatorname{argmin}_{\mathbf{y} \in \mathcal{Y}} h(\mathbf{x}_k, \mathbf{y}).$

Schatten-p norm: $\|\mathbf{X}\|_{S_p} = \left(\sum_i \sigma_i^p(\mathbf{X}) \right)^{1/p}$, low-rankness regularizer.

Theorem 1. *With compatiable dimensions and* $\frac{1}{p} = \sum_{i=1}^{I} \frac{1}{p_i}$:

$$\frac{1}{p} \|\mathbf{X}\|_{S_p}^p = \min_{\mathbf{X} = \Pi_{i=1}^I \mathbf{X}_i} \sum_{i=1}^{I} \frac{1}{p_i} \|\mathbf{X}_i\|_{S_{p_i}}^{p_i}.$$

If $0 < p < 1$, we can still choose $p_i \geq 1$.

Majorization Minimization

- How to choose the majorant function?

$$|x|^p = \min_{a>0} \frac{1}{c}(x^2/a + a^q),$$

half-quadratic

$$\text{where } q = \frac{p}{2-p}, c = q^{1/(q+1)} + q^{-q/(q+1)}, a^* = (x^2/q)^{1/(q+1)}.$$

l_p-norm, sparsity regularizer

$$\|\mathbf{x}\|_p^p \Longrightarrow \frac{1}{c}(\mathbf{x}^T \mathbf{A}_k^{-1} \mathbf{x} + \mathbf{1}^T \mathbf{A}_k^q \mathbf{1}) \Longrightarrow c^{-1} \mathbf{x}^T \mathbf{A}_k^{-1} \mathbf{x}$$

$$\Longrightarrow \frac{q}{q+1} \mathbf{x}^T \operatorname{Diag}(|\mathbf{x}_k|^{-2/(q+1)}) \mathbf{x} \Longrightarrow \frac{p}{2} \mathbf{x}^T \operatorname{Diag}(|\mathbf{x}_k|)^{p-2} \mathbf{x},$$

$$\text{where } \mathbf{A}_k = \operatorname{Diag}(\mathbf{a}_k), a_{k,i} = (x_{k,i}^2/q)^{1/(q+1)}.$$

Iteratively Reweighted Least Squares (IRLS)

Majorization Minimization

- How to choose the majorant function?

$$\|\mathbf{X}\|_{S_p}^p = \|\boldsymbol{\sigma}(\mathbf{X})\|_p^p \Longrightarrow \frac{p}{2}\boldsymbol{\sigma}(\mathbf{X})^T \operatorname{Diag}(\boldsymbol{\sigma}(\mathbf{X}_k)^{p-2})\boldsymbol{\sigma}(\mathbf{X})$$

$$= \frac{p}{2}\operatorname{tr}(\operatorname{Diag}(\boldsymbol{\sigma}(\mathbf{X}))\operatorname{Diag}(\boldsymbol{\sigma}(\mathbf{X}_k)^{p-2})\operatorname{Diag}(\boldsymbol{\sigma}(\mathbf{X})))$$

$$= \frac{p}{2}\operatorname{tr}(\mathbf{X}^T(\mathbf{X}_k\mathbf{X}_k^T)^{p/2-1}\mathbf{X}).$$

Schatten-p norm, low-rankness regularizer

$$\operatorname{tr}\left((\mathbf{X}\mathbf{X}^T)^{p/2}\right) \Longrightarrow \operatorname{tr}\left((\mathbf{X}_k\mathbf{X}_k^T)^{p/2-1}\mathbf{X}\mathbf{X}^T\right) \Longrightarrow \operatorname{tr}\left(\mathbf{X}^T(\mathbf{X}_k\mathbf{X}_k^T)^{p/2-1}\mathbf{X}\right).$$

$$\operatorname{tr}\left((\mathbf{X}\mathbf{X}^T)^{p/2}\right) = \min_{\mathbf{A}\succ\mathbf{0}} \frac{1}{c}[\operatorname{tr}(\mathbf{A}^{-1}\mathbf{X}\mathbf{X}^T) + \operatorname{tr}(\mathbf{A}^q)].$$

$$\operatorname{tr}(\mathbf{A}^{-1}\mathbf{X}\mathbf{X}^T) + \operatorname{tr}(\mathbf{A}^q) \geq \sum_{i=1}^n \sigma_i(\mathbf{X}\mathbf{X}^T)\sigma_{n-i+1}(\mathbf{A}^{-1}) + \sum_{i=1}^n \sigma_i^q(\mathbf{A})$$

$$= \sum_{i=1}^n \frac{\sigma_i^2(\mathbf{X})}{\sigma_i(\mathbf{A})} + \sum_{i=1}^n \sigma_i^q(\mathbf{A})$$