Convex-Composite Optimization
Convex-Composite Model

We now consider problems of the form

$$\min f(x) := h(F(x))$$

where \( h : \mathbb{E} \to \overline{\mathbb{R}} \) is a closed proper convex function and \( F : \mathbb{E} \to \mathbb{Y} \) is continuously differentiable.

In general, the functions \( h \circ F \) are neither differentiable or convex. However, the nonsmoothness is of a familiar form since it arises from the convex function \( h \).
Convex-Composite Model

We now consider problems of the form

$$\min f(x) := h(F(x))$$

where $h : \mathbb{E} \rightarrow \overline{\mathbb{R}}$ is a closed proper convex function and $F : \mathbb{E} \rightarrow \mathbb{Y}$ is continuously differentiable.

In general, the functions $h \circ F$ are neither differentiable or convex. However, the nonsmoothness is of a familiar form since it arises from the convex function $h$.

Most problems from nonlinear programming can be cast in this framework.
Nonlinear least squares

Let $F : \mathbf{E} \to \mathbf{Y}$ with $m = \dim \mathbf{Y} >> \dim \mathbf{E} = n$ and consider the equation $F(x) = 0$.

Since $m > n$ it is highly unlikely that a solution to this equation exists. However, one might try to obtain a best approximate solution by solving the problem

$$\min \{ \|F(x)\| : x \in \mathbf{E}\}.$$

This is a convex composite optimization problem since the norm is a convex function.
Nonlinear convex inclusions

Let $F : \mathbf{E} \to \mathbf{Y}$ with $m = \dim Y >> \dim \mathbf{E} = n$ and consider the inclusion $F(x) \in C$ where $C \subset \mathbf{Y}$ is nonempty closed cvx.

Since $m > n$ it is again highly unlikely that a solution to this equation exists. However, one might try to obtain a best approximate solution by solving the problem

$$\min \{ \text{dist} (F(x) | C) : x \in \mathbf{E} \}. $$

This is a convex composite optimization problem since the distance to a convex set is cvx.

The set $C$ is often a cone such as $\mathbf{S}^n_+$ or $\mathbf{R}^k \times \{0\}^{m-k}$. 
Nonlinear Programming (NLP)

Let $F : E \to Y$, $C \subset Y$ a non-empty closed convex set, and $f_0 : E \to \mathbb{R}$, and consider the constrained optimization problem

$$\min\{ f_0(x) : F(x) \in C \} = \min f_0(x) + \delta_C(F(x)).$$

This is a convex composite optimization problem since $h(\mu, y) := \mu + \delta_C(y)$ is cvx.
Exact Penalization

Again consider the NLP

$$\min \{ f_0(x) | F(x) \in C \} = \min f_0(x) + \delta_C(F(x)).$$

One can approximate this problem by the unconstrained optimization problem

$$\min \{ f_0(x) + \alpha \text{dist} (f(x) | C) : x \in E \}.$$ 

This is a convex composite optimization problem where

$$h(\eta, y) = \eta + \alpha \text{dist} (y | C)$$

is a convex function.
Exact Penalization

Again consider the NLP

\[
\min \{ f_0(x) \mid F(x) \in C \} = \min f_0(x) + \delta_C(F(x)).
\]

One can approximate this problem by the unconstrained optimization problem

\[
\min \{ f_0(x) + \alpha \text{dist} (f(x) \mid C) : x \in E \}.
\]

This is a convex composite optimization problem where

\[ h(\eta, y) = \eta + \alpha \text{dist} (y \mid C) \]

is a convex function.

The function \( f_0(x) + \alpha \text{dist} (f(x) \mid C) \) is called an exact penalty function for the problem \( \min \{ f_0(x) : F(x) \in C \} \).
First-Order theory for CVX-Comp

Consider the cvx-comp objective $h \circ F$. If $h$ is finite-valued, we know it is locally Lipschitz. Consequently,

$$f(y) = h(F(y)) = h(F(x) + F'(x)(y - x)) + o(\|y - x\|).$$
First-Order theory for CVX-Comp

Consider the cvx-comp objective $h \circ F$. If $h$ is finite-valued, we know it is locally Lipschitz. Consequently,

$$f(y) = h(F(y)) = h(F(x) + F'(x)(y - x)) + o(\|y - x\|).$$

Given $d \in \mathbf{E}$, we can rewrite this equation as

$$h(F(x + d)) = h(F(x)) + \Delta f(x; d) + o(\|d\|) \quad \text{where}$$

$$\Delta f(x; d) := h(F(x) + F'(x)d) - h(F(x)).$$
First-Order theory for CVX-Comp

Consider the cvx-comp objective \( h \circ F \). If \( h \) is finite-valued, we know it is locally Lipschitz. Consequently,

\[
f(y) = h(F(y)) = h(F(x) + F'(x)(y - x)) + o(\|y - x\|).
\]

Given \( d \in \mathbf{E} \), we can rewrite this equation as

\[
h(F(x + d)) = h(F(x)) + \Delta f(x; d) + o(\|d\|) \quad \text{where} \quad \Delta f(x; d) := h(F(x) + F'(x)d) - h(F(x)).
\]

Then, for every \( d \in \mathbf{E} \),

\[
f'(x; d) = \lim_{t \downarrow 0} \frac{f(x + td) - f(x)}{t}
= \lim_{t \downarrow 0} \frac{\Delta f(x; td)}{t} + \frac{o(t)}{t}
= h'(F(x); F'(x)d).
\]

That is, \( f \) is directionally differentiable on \( \mathbf{E} \) in all directions.
Recall the notion of \textit{regular} subdifferential defined earlier for potentially non-convex functions:

\[
\hat{\partial} f(x) := \{ v \mid f(x) + \langle v, y - x \rangle \leq f(y) + o(\|y - x\|) \quad \forall y \in \mathbf{E} \}.
\]

We showed that \( \hat{\partial} f(x) \) is a closed convex set that coincides with \( \partial f(x) \) when \( f \) is convex.
Recall the notion of regular subdifferential defined earlier for potentially non-convex functions:

\[ \hat{\partial} f(x) := \{ v \mid f(x) + \langle v, y - x \rangle \leq f(y) + o(\|y - x\|) \quad \forall y \in \mathbf{E} \}. \]

We showed that \( \hat{\partial} f(x) \) is a closed convex set that coincides with \( \partial f(x) \) when \( f \) is convex.

When \( f \) is cvx-comp, for every \( v \in \hat{\partial} f(x) \), we have

\[ \langle v, d \rangle \leq \frac{f(x + td) - f(x)}{t} = \frac{\Delta f(x; td)}{t} + \frac{o(t)}{t} \quad \forall t > 0. \]

Hence

\[ \langle v, d \rangle \leq h'(F(x); F'(x)d) = \delta^*(F'(x)d | \partial h(F(x))) = \delta^*(d | F'(x) \partial h(F(x))). \]

So that

\[ \delta^*(d | \hat{\partial} f(x)) \leq \delta^*(d | F'(x) \partial h(F(x))) \quad \implies \quad \hat{\partial} f(x) \subset F'(x) \partial h(F(x)). \]
\[ \partial f(x) = F'(x)^* \partial h(F(x)) \]

On the other hand, we have

\[
\begin{align*}
f(y) &= h(F(x) + F'(x)(y - x)) + o(\|y - x\|) \\
&\geq h(F(x)) + \langle v, F'(x)(y - x) \rangle + o(\|y - x\|) \quad \forall \ v \in \partial h(F(x)) \\
&= f(x) + \langle F'(x)^* v, (y - x) \rangle + o(\|y - x\|) \quad \forall \ v \in \partial h(F(x)).
\end{align*}
\]

Hence,

\[ F'(x)^* \partial h(F(x)) \subset \hat{\partial} f(x). \]

Consequently,

\[ \hat{\partial} f(x) = F'(x)^* \partial h(F(x)) \quad \text{and} \quad f'(x; d) = \delta^*(d| \hat{\partial} f(x)). \]

For this reason, when \( f \) is finite-valued cvx-comp, we write \( \partial f(x) \) instead of \( \hat{\partial} f(x) \) and call \( \partial f(x) \) the subdifferential of \( f \) at \( x \).
Directional Derivative Approximation

In our development of numerical methods for minimizing convex composite functions, we make extensive use of the difference function

$$\Delta f(x; d) := h(F(x) + F'(x)d) - h(F(x)).$$

In particular, it is often used as a surrogate for the directional derivative $f'(x; d)$. In this respect, recall that

$$\lambda_1^{-1} \Delta f(x; \lambda_1 d) \leq \lambda_2^{-1} \Delta f(x; \lambda_2 d) \quad \text{for } 0 < \lambda_1 \leq \lambda_2,$$

due to the non-decreasing nature of the difference quotients. An important consequence of this inequality is that

$$f'(x; d) = \inf_{t > 0} t^{-1} \Delta f(x; td) \leq \Delta f(x; d),$$

which also implies that

$$\Delta f(x; td) \leq t \Delta f(x; d) \quad \forall t > 0.$$
Theorem: Let $h : Y \to \mathbb{R}$ be convex and $F : E \to Y$ be continuously differentiable. If $\bar{x}$ is a local solution to the problem $\min \{ h(F(x)) \}$, then $0 \in \partial f(\bar{x})$. Moreover, the following conditions are equivalent:

(a) $0 \in \partial f(x)$.

(b) $d = 0$ is a global solution to $\min_{d \in E} h(F(\bar{x}) + F'(\bar{x})d)$.

(c) $0 \leq h'(F(x); F'(x)d)$ for all $d \in E$.

(d) $0 \leq \Delta f(x; d)$ for all $d \in E$. 

Optimality Conditions for Cvx Comp Optimization
Proof: Let $\bar{x}$ be a local solution to $\min\{h(F(x))\}$ and set $\Psi(d) := h(F(\bar{x}) + F'(\bar{x})d)$. Then $0 \leq f'(\bar{x}; d)$ for all $d \in E$. Since $f'(\bar{x}; \cdot) = \delta^*_{\partial f(\bar{x})}$, it must be the case that $0 \in \partial f(x)$.

[(a) $\iff$ (b)] Since $\Psi$ is convex and $\partial \Psi(0) = F'(\bar{x})^* \partial h(F(\bar{x})) = \partial f(\bar{x})$, we have $0 \in \partial \Psi(0)$ so $d = 0$ is a global solution to $\min_d \Psi(d)$.

[(a) $\iff$ (c)] This follows from the fact that $f'(\bar{x}; d) = h'(F(x); F'(\bar{x})d)$.

[(c) $\implies$ (d)] Due to the convexity of $\Psi$, $h'(F(x); F'(\bar{x})d) \leq \Delta f(x; d)$ for all $d \in E$ so (c) implies (d).

[(d) $\implies$ (b)] (d) implies that $h(F(\bar{x})) \leq h(F(\bar{x}) + F'(\bar{x})d)$ for all $d \in E$ so that (b) holds.
Line–Search Methods

Let \( f : \mathbb{E} \to \mathbb{R} \) and consider the problem \( \min_x f(x) \).

We consider iterative schemes of the form

\[
x_{k+1} := x_k + \lambda_k d_k,
\]

where it is intended that \( f(x_{k+1}) < f(x_k) \).

Such methods are called descent methods. The scalar \( \lambda_k > 0 \) is called the \textit{step length} and the vector \( d_k \) is called the \textit{search direction}.

Observe that

\[
\{ d : f'(x; d) < 0 \} \subset \{ d : \exists \bar{\lambda} > 0, \text{ s.t. } f(x + \lambda d) < f(x) \forall \lambda \in (0, \bar{\lambda}) \}.
\]

Thus, one way to achieve descent is to choose the search direction from the set \( \{ d : f'(x_0; d) < 0 \} \).
Cauchy and Gauss-Newton search directions

The search direction $d_k$ obtained by solving

$$\min \{ f'(x_k; d) : \|d\| \leq 1 \}.$$

is called the direction of steepest descent, or the Cauchy direction.

The search direction $d_k$ obtained by solving

$$\min_{\|d\| \leq \beta} \Delta f(x_k; d) + \frac{1}{2\alpha} \|d\|^2$$

is called the prox-Newton or Gauss-Newton search direction. Here $0 < \alpha, \beta \leq \infty$ with infinite values allowed.
The Backtracking line search

Consider the finite-valued cvx-comp framework $f = h \circ F$. Let $c, \gamma \in (0, 1)$ and let $x_k, d_k \in E$ be such that $\Delta f(x_k; d) < 0$.

**Backtracking Line Search:**

$$
\lambda_k := \max \gamma^s
$$

subject to $s \in \{0, 1, 2, \ldots\}$ and

$$
h(F(x + \gamma^s d)) \leq h(F(x)) + c\gamma^s \Delta f(x_kd_k).
$$

The value $\lambda_k$ is called the backtracking step size.
Backtracking Descent Algorithm

**Algorithm:** Backtracking Descent

**Input:** Initial point \( x_0 \in \mathbb{E} \) and line search parameters \( c, \gamma \in (0, 1) \).

**For:** \( k = 1, 2, \ldots \)

**Search Direction:** Let \( D_k \subset \{ d : \Delta f(x_k; d) < 0 \} \). If \( D_k = \emptyset \) stop; otherwise choose \( d_k \in D_k \).

**Backtracking line search:**

\[ \lambda_k := \max \gamma^s \]

subject to \( s \in \{0, 1, 2, \ldots \} \) and

\[ h(F(x + \gamma^s d)) \leq h(F(x)) + c\gamma^s \Delta f(x_k d_k). \]

**Update:** Set \( x_{k+1} := x_k + \lambda_k d_k \) and \( k := k + 1 \).
Convergence of Backtracking Descent Algorithm

**Theorem:** Let $f : \mathbb{E} \to \mathbb{R}$ be given by $f(x) = h(F(x))$ where $h : \mathbb{Y} \to \mathbb{R}$ is convex and $F : \mathbb{E} \to \mathbb{Y}$ is differentiable. Let $x_0 \in \mathbb{R}^n$ and assume that

(a) $h$ is Lip. cont. on the set $\{y : h(y) \leq h(F(x_0))\}$, and

(b) $F'$ is uniformly continuous on the set $\overline{co}\{x : h(F(x)) \leq h(F(x_0))\}$.

If $\{x_k\}$ is the sequence generated by the algorithm initiated at $x_0$, then one of the following must occur:

(i) There is a $k_0$ such that $D_{k_0} = \emptyset$.

(ii) $f(x_k) \downarrow -\infty$.

(iii) The sequence $\{\|d_k\|\}$ diverges to $+\infty$.

(iv) For every subsequence $J \subset \mathbb{N}$ for which $\{d_k\}_J$ is bounded, we have

$$\lim_{J} \Delta f(x_k; d_k) = 0.$$
Proof: Spps to the contrary that none of (i) – (iv) occur. Then \( \exists J \subset \mathbb{N} \) such that \( \{d_j\}_J \) is bounded and there is a \( \beta > 0 \) with 
\[
\sup_J \Delta f(x_j; d_j) \leq -\beta < 0.
\]

Since \( \{f(x_j)\} \) is a decr. seq. that is bounded below, \( f(x_j) \to f^* \) for some \( f^* \in \mathbb{R} \). Consequently, \( (f(x_{j+1}) - f(x_j)) \to 0 \).
Convergence of Backtracking Descent Algorithm

**Proof:** Spps to the contrary that none of (i) – (iv) occur. Then \( \exists J \subset \mathbb{N} \) such that \( \{d_j\}_J \) is bounded and there is a \( \beta > 0 \) with
\[
\sup_J \Delta f(x_j; d_j) \leq -\beta < 0.
\]
Since \( \{f(x_j)\} \) is a decr. seq. that is bounded below, \( f(x_j) \to f^* \) for some \( f^* \in \mathbb{R} \). Consequently, \( (f(x_{j+1}) - f(x_j)) \to 0 \).

The choice of \( \lambda_k \) implies that \( \lambda_j \Delta f(x_j; d_j) \to 0 \). Therefore, \( \lambda_j \to 0 \) so WLOG \( \lambda_j < 1 \) for all \( j \in J \). Again, the choice of \( \lambda_j \) implies that
\[
c\lambda_j \gamma^{-1} \Delta f(x_j; d_j) \leq f(x_j + \lambda_j \gamma^{-1} d_j) - f(x_j) \quad \forall j \in J.
\]
Convergence of Backtracking Descent Algorithm

**Proof:** Spps to the contrary that none of (i) – (iv) occur. Then \( \exists J \subset \mathbb{N} \) such that \( \{d_j\}_J \) is bounded and there is a \( \beta > 0 \) with
\[
\sup_J \Delta f(x_j; d_j) \leq -\beta < 0.
\]
Since \( \{f(x_j)\} \) is a decr. seq. that is bounded below, \( f(x_j) \to f^* \) for some \( f^* \in \mathbb{R} \). Consequently, \( (f(x_{j+1}) - f(x_j)) \to 0 \).

The choice of \( \lambda_k \) implies that \( \lambda_j \Delta f(x_j; d_j) \to 0 \). Therefore, \( \lambda_j \to 0 \) so WLOG \( \lambda_j < 1 \) for all \( j \in J \). Again, the choice of \( \lambda_j \) implies that
\[
c\lambda_j \gamma^{-1} \Delta f(x_j; d_j) \leq f(x_j + \lambda_j \gamma^{-1} d_j) - f(x_j) \quad \forall j \in J.
\]
But,
\[
f(x_j + \lambda_j \gamma^{-1} d_j) - f(x_j) \leq \lambda_j \gamma^{-1} \Delta f(x_j; d_j) + K \|F(x_j + \lambda_j \gamma^{-1} d_j) - (F(x_j) + \lambda_j \gamma^{-1} F'(x_j) d_j)\|
\]
\[
\leq \lambda_j \gamma^{-1} \Delta f(x_j; d_j) + K \lambda_j \gamma^{-1} \|d_j\| \int_0^1 \|F'(x_j + \tau \gamma^{-1} \lambda_j d_j) - F'(x_j)\| d\tau
\]
\[
\leq \lambda_j \gamma^{-1} \{\Delta f(x_j; d_j) + K \|d_j\| \omega(\gamma^{-1} \lambda_j \|d_j\|)\}
\]
for all \( j \in J \), where \( K \) is a Lipschitz constant for \( h \) and \( \omega \) is the modulus of continuity for \( F' \).
Convergence of Backtracking Descent Algorithm

**Proof:** Spps to the contrary that none of (i) – (iv) occur. Then \( \exists J \subset \mathbb{N} \) such that \( \{d_j\}_J \) is bounded and there is a \( \beta > 0 \) with

\[
\sup_J \Delta f(x_j; d_j) \leq -\beta < 0.
\]

Since \( \{f(x_j)\} \) is a decr. seq. that is bounded below, \( f(x_j) \to f^* \) for some \( f^* \in \mathbb{R} \). Consequently, \( (f(x_{j+1}) - f(x_j)) \to 0 \).

The choice of \( \lambda_k \) implies that \( \lambda_j \Delta f(x_j; d_j) \to 0 \). Therefore, \( \lambda_j \to 0 \) so WLOG \( \lambda_j < 1 \) for all \( j \in J \). Again, the choice of \( \lambda_j \) implies that

\[
c \lambda_j \gamma^{-1} \Delta f(x_j; d_j) \leq f(x_j + \lambda_j \gamma^{-1} d_j) - f(x_j) \quad \forall j \in J.
\]

But,

\[
f(x_j + \lambda_j \gamma^{-1} d_j) - f(x_j) \\
\leq \lambda_j \gamma^{-1} \Delta f(x_j; d_j) + K \|F(x_j + \lambda_j \gamma^{-1} d_j) - (F(x_j) + \lambda_j \gamma^{-1} F'(x_j) d_j)\| \\
\leq \lambda_j \gamma^{-1} \Delta f(x_j; d_j) + K \lambda_j \gamma^{-1} \|d_j\| \int_0^1 \|F'(x_j + \tau \gamma^{-1} \lambda_j d_j) - F'(x_j)\|d\tau \\
\leq \lambda_j \gamma^{-1} \{\Delta f(x_j; d_j) + K \|d_j\| \omega(\gamma^{-1} \lambda_j \|d_j\|)\}
\]

for all \( j \in J \), where \( K \) is a Lipschitz constant for \( h \) and \( \omega \) is the modulus of continuity for \( F' \).

Therefore,

\[
0 < (1 - c) \Delta f(x_j; d_j) + K \omega(\lambda_j \gamma^{-1} \|d_j\|) \|d_j\| \\
\leq (c - 1) \beta + K \omega(\lambda_j \gamma^{-1} \|d_j\|) \|d_j\|
\]

for all \( j \in J \). Letting \( j \in J \) go to \( \infty \), we obtain the contradiction \( 0 \leq (c - 1) \beta < 0 \).
Corollary: Let $f$ and $\{x_k\}$ be as in the statement of Theorem and let $\tau \in (0, 1)$ and $\{\delta_k\} \subset (\delta, \bar{\delta})$ for some $\bar{\delta} \geq \delta > 0$. Suppose that

(a) $f$ is bounded below, and

(b) $D_k := \{d \in \delta_k \mathbb{B} | \Delta f(x_k; d) \leq \tau \Delta_k f(x_k)\}$, where

$$\Delta_k f(x_k) := \min \{\Delta f(x_k; d) | \|d\| \leq \delta_k\}.$$ 

Then every cluster, $\bar{x}$, point of the sequence $\{x_j\}$ satisfies $0 \in \partial f(\bar{x})$. 

Convergence of Backtracking Descent Algorithm
Convergence of Backtracking Descent Algorithm

**Proof:** By the Theorem, $\Delta f(x_j; d_j) \to 0 \implies \Delta_k f(x_k) \to 0$. For $j \in \mathbb{N}$, let $bd_j \in \operatorname{argmin} \{\Delta f(x_k; d) \mid \|d\| \leq \delta_k\}$. If $J \subset \mathbb{N}$ is such that $x_j \rightarrow \bar{x}$ we can always refine $J$ if necessary to get that $(d_j, \bar{d}_j, \delta_j) \rightarrow (\bar{d}, \tilde{d}_j, \tilde{\delta})$ for some $\bar{d}, \tilde{d} \in \delta \mathbb{B}$ and $\tilde{\delta} \in (\delta, \bar{\delta})$. But then $\Delta f(\bar{x}; \bar{d}) = \Delta f(\bar{x}; \tilde{d}) = 0$ which implies that

$$h(F(\bar{x}) + F'(\bar{x})\bar{d}) = h(F(\bar{x}) + F'(\bar{x})\tilde{d}) = h(F(\bar{x})).$$

Note that

$$h(F(x_j) + F'(x_j)\bar{d}_j) \leq h(F(x_j) + F'(x_j)d) \quad \forall d \in \delta_j \mathbb{B}.$$ 

Hence, in the limit over $J$,

$$h(F(\bar{x}) + F'(\bar{x})\tilde{d}) \leq h(F(\bar{x}) + F'(\bar{x})d) \quad \forall d \in \tilde{\delta} \mathbb{B}.$$
Convergence of Backtracking Descent Algorithm

Consequently,

\[ \tilde{d} \in \arg\min \{ h(F(\bar{x}) + F'(\bar{x})d) : \|d\| \leq \tilde{\delta} \}. \]

But \( h(F(\bar{x})) = h(F(\bar{x}) + F'(\bar{x})\tilde{d}) \) so that
\[ 0 \in \arg\min \{ h(F(\bar{x}) + F'(\bar{x})d) : \|d\| \leq \tilde{\delta} \}. \]

Since \( h(F(\bar{x}) + F'(\bar{x})d) \) is convex, \( d = 0 \) is a global solution to the problem \( \min \{ h(F(\bar{x}) + F'(\bar{x})d) \} \). Therefore, by the optimality condition theorem,

\[ 0 \in \partial f(\bar{x}). \]