Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 247–253
ISBN 978-83-60810-22-4

Improved asymptotic analysis for SUMT methods

Jean-Pierre Dussault
Département d'informatique
Faculté des Sciences
Université de Sherbrooke
Sherbrooke (Québec) CANADA J1K 2R1
Email: Jean-Pierre.Dussault@USherbrooke.ca

***Abstract*—We consider the SUMT (*Sequential Unconstrained Minimization Technique*) method using extrapolations to link successive unconstrained sub-problems. The case when the extrapolation is obtained by a first order Taylor estimate and Newton's method is used as a correction in this predictor-corrector scheme was analyzed in [1]. It yields a two–steps super-linear asymptotic convergence with limiting order of $\frac{4}{3}$ for the logarithmic barrier and order two for the quadratic loss penalty.**

We explore both lower order variants (approximate extrapolations correction computations) as well as higher order variants (second order and further) Taylor estimate.

First, we address inexact solutions of the linear systems arising within the extrapolation and the Newton's correction steps. Depending on the inexactness allowed, asymptotic convergence order reduces, more severely so for interior variants.

Second, we investigate the use of higher order path following strategies in those methods. We consider the approach based on a high order expansion of the so-called central path, somewhat reminiscent of Chebyshev's third order method and its generalizations. The use of higher order representation of the path yields spectacular improvement in the convergence property, even more so for the interior variants.

I. Introduction

We consider non linear programs (NLP) of the form

$$\begin{aligned} &\min_{x\in\mathbb{R}^n} f(x) \\ &\text{subject to } g(x) \le 0 \end{aligned} \tag{1}$$

or

$$\begin{aligned} &\min_{x\in\mathbb{R}^n} f(x) \\ &\text{subject to } g(x) = 0 \end{aligned} \tag{2}$$

with $g : \mathbb{R}^n \to \mathbb{R}^m$.

We will address both formulations using SUMT, Eq.(1) using the classical log barrier method and Eq.(2) will be developed using the quadratic loss penalty function. A slight emphasis is put on the logarithmic barrier variant.

Fiacco and McCormick [5] pioneered the study of SUMT, and obtained the important result that close to a solution, the unconstrained sub-problems induce differential trajectories, and proposed the use of extrapolation to follow the trajectories. In linear programming, Mehrotra [6] popularized the use of so called predictor-corrector algorithms, intimately related to the extrapolations of the SUMT trajectory.

This research was partially supported by NSERC grant OGP0005491

We present asymptotic results related to inexact versions of both variants, looking for super-linear convergence of order lower than 2, as well as high order extrapolations to aim for faster asymptotic convergence.

Results from [1] state that interior variants achieve a limiting convergence order of two–steps $\frac{4}{3}$ while exterior variants reach a limiting 2–steps quadratic order. We call two–steps convergence since the convergence order requires the solution of two linear systems, one to compute the extrapolation and another one to compute the Newton step.

Asymptotic order is not all, and interior variants are known to yield polynomial complexity in a wide variety of contexts while exterior variants could be plagued by combinatorial aspects related to active set identification when applied to inequality constrained problems.

We will first analyze inexact versions of SUMT, and conclude that the deterioration with respect to the asymptotic order is more severe for interior variants, exterior variants' degradation being benign.

On the other hand, high order versions bring both variants comparable with respect to their asymptotic behavior. For example, the second order extrapolation allows to bypass any Newton correction asymptotically. The second order correction involves solving two linear systems, but both systems share the same matrix. Therefore, only one factorization is require. For unstructured dense problems, factorisation of a matrix in a linear system entails $\mathcal{O}(n^3)$ arithmetic operations while solving the system needs a further $\mathcal{O}(n^2)$ operations. The second order extrapolation requires only one factorization, much better than two independant linear systems.

Finally, we propose an approximate high order version based on the Shamanskii approach which is simple to implement and shares the good asymptotic improvements of the exact high order versions.

II. SUMT basic extrapolations

We now recall two path following approaches, allowing to settle our notation and present some basic properties. Path following methods are related to the so-called central path, and involves steps named as predictor, corrector, centrality corrector, higher order predictor. We will detail below the terms we will use.

We stress that the results for the exterior quadratic loss penalty function and the interior log-barrier function are very similar. The only difference is the limiting convergence order, $\frac{4}{3}$ for the interior approach and 2 for the exterior penalty.

A. Log–Barrier

The log barrier approach to solve Eq.(1) consists in solving a sequence of sub-problems of the form

$$\min_{g(x)<0} \phi(x, \rho_k) = f(x) - \rho_k \sum \log(-g_i(x)) \tag{3}$$

in the interior of the feasible set $E = \{x : g(x) \le 0\}$. Writing out the optimality conditions

$$\nabla_x \phi(x, \rho_k) = \nabla f(x) - \rho_k \sum \frac{1}{g_i(x)} \nabla g_i(x) = 0 \tag{4}$$

and by making the substitution $y_i = \frac{\rho_k}{g_i(x)}$ and introducing the residual Φ_k, one arrives at the primal-dual equations

$$\Theta(x, y, \Phi_k, \rho_k) = \begin{cases} \nabla f(x) - y\nabla g(x) = \Phi_k \\ \quad\quad\quad yG(x) = \rho_k e \end{cases} \tag{5}$$

where $e = (1, 1 \ldots 1)^t$ and $G(x) = diag(g_i(x))$.

Under suitable assumptions, this last system of equations implicitly defines differentiable trajectories $x(\rho, \Phi)$ and $y(\rho, \Phi)$ close to $\rho = 0$.

B. Quadratic loss

The quadratic loss approach to solve Eq.(2) consists in solving a sequence of sub-problems of the form

$$\min \phi(x, \rho_k) = f(x) + \frac{1}{\rho_k} \|g(x)\|^2. \tag{6}$$

Writing out the optimality conditions

$$\nabla_x \phi(x, \rho_k) = \nabla f(x) + \frac{g(x)^t}{\rho_k} \nabla g(x) = 0 \tag{7}$$

and by making the substitution $y = \frac{g(x)^t}{\rho_k}$ and introducing the residual Φ_k, one arrives at the primal-dual equations

$$\Theta(x, y, \Phi_k, \rho_k) = \begin{cases} \nabla f(x) + y\nabla g(x) = \Phi_k \\ \quad\quad\quad g(x) = \rho_k y \end{cases} \tag{8}$$

Under suitable assumptions, this last system of equations implicitly defines differentiable trajectories $x(\rho, \Phi)$ and $y(\rho, \Phi)$ close to $\rho = 0$.

C. Common properties

Penalty and barriers trajectories share much properties. Those may be expressed conveniently using the Θ function in a unified way. In the following result, g_{I^*} refers to the active constraints in the log barrier case, and the whole g vector in the quadratic penalty case.

Theorem 2.1: [1]Let x^* be a regular point of the constraints $g_{I^*}(x) = 0$ which satisfies to the second order sufficient optimality conditions for (1) as well as to the strict complementarity condition $y_{I^*} > 0$ for the log barrier case. If the functions f and g are $\mathcal{C}^p(\mathbb{R}^n)$, then there exists differentiable trajectories $x(\rho, \Phi)$ and $y(\rho, \Phi)$ of class $\mathcal{C}^{p-1}(\mathbb{R}^n)$ such that

1) $x(0, 0) = x^*$ and $y(0, 0) = y^*$;
2) if ρ and $\|\Phi\|$ are small enough, $x(\rho, 0)$ satisfies to the second order sufficient optimality conditions for the penalized sub-problems $\min f(x) - \rho \sum_{i=1}^m \log(-g_i(x))$, where $x(r, \Phi)$, $y(r, \Phi)$ are solutions of the following equations:

$$\Theta(x, y, \Phi, 0) = 0 \tag{9}$$

Moreover, the following bounds hold asymptotically:

a) $\|x(\rho, \Phi) - x^*\| \sim \mathcal{O}(\max(\rho, \|\Phi\|))$;
b) $\|y(\rho, \Phi) - y^*\| \sim \mathcal{O}(\max(\rho, \|\Phi\|))$;
c) $\|g_{I^*}(x(\rho, \Phi))\| \sim \mathcal{O}(\rho)$.

Remark 2.1: Although we use primal-dual equations, in this SUMT variant, the dual variables y are dependent on the primal x, so that global convergence in infered from the fact that the penalty or barrier is minimized (using globally convergent algorithms), allowing to prove that cluster points of the generated sequence are indeed stationary.

We will denote $G(x) = diag(g_i(x))$ and for the log barrier,

$$\begin{aligned} \Phi(x, \rho) &= \nabla_x \phi(x, \rho) && (10) \\ &= \nabla f(x) - \sum \frac{\rho}{g_i(x)} \nabla g_i(x) && (11) \\ &= \nabla f(x) - \rho \nabla g^t(x) G(x)^{-1} e. && (12) \end{aligned}$$

ϕ is closely related to the Lagrangian $l(x, \lambda) = f(x) + g(x)\lambda$, and by defining $\lambda = -\rho G(x)^{-1} e$, i.e. $\lambda_i = -\frac{\rho}{g_i(x)}$, $\nabla_x l(x, \lambda) = L(x, \lambda) = \nabla f(x) + \lambda \nabla g(x) = \Phi(x, \rho)$.

Similarly, for the quadratic penalty, $\lambda = \frac{g(x)^t}{\rho}$ and $\Phi(x, \rho) = \nabla f(x) + \frac{g(x)^t}{\rho} \nabla g(x)$.

We are concerned with approximate solutions $x(\rho, r)$ which satisfy $\Phi(x(\rho, r), \rho) = r$. In the sequel the residual r is assumed to satisfy $\|r\| \sim \rho$.

The basic predictor-corrector path following approach consists then in having an estimate $x(\rho, r)$ which satisfy $\Phi(x(\rho, r), \rho) = r$ and then iterate the following two steps:

pred	extrapolate $\hat{x}^1 = x + \frac{\partial x}{\partial \rho}(\rho^+ - \rho) + \frac{\partial x}{\partial r}(-r)$
corr	perform Newton corrections from $\hat{x}^1$ on the problem Eq.(6) or Eq.(3) for ρ^+ until $\|\Phi(x, \rho^+)\| \le \rho^+$.

For this basic scheme and the log barrier case, a single Newton correction asymptotically yields $x(\rho^+, r^+)$ with $\|r^+\| \le \rho^+$ provided that $\frac{\rho^+}{\rho^{\frac{4}{3}}} \to 0$ yielding a two–steps superlinear convergence of limiting order $\frac{4}{3}$. If one is prepared to perform two Newton corrections, then the limiting order is improved to $\frac{\rho^+}{\rho^{\frac{8}{5}}} \to 0$ [4] yielding a three-steps super-linear convergence of limiting order $\frac{8}{5}$. Using a measure similar to Ostrovski efficiency, the optimal strategy for this family of algorithms is to aim for two Newton corrections following an extrapolation [4].

For the quadratic loss case, a single Newton correction asymptotically yields $x(\rho^+, r^+)$ with $\|r^+\| \le \rho^+$ provided that $\frac{\rho^+}{\rho^2} \to 0$ yielding a two–steps super-linear convergence of limiting quadratic order.

D. Terminology

First order extrapolations are usually named "predictor" steps. Higher order terms are sometimes named "corrections" to the predictor, but we will stick to the terminology "higher order". Once a predictor (of arbitrary order) is computed, sometimes it is necessary to perform Newton iterations, refered to as a "corrector" steps. This is sometimes named "centrality correction" steps.

Predictor steps aim at changing the trajectory parameter ρ to a smaller value while corrector steps aim at improving the parametric solution for a given ρ-value.

III. Inexact versions

Since extrapolations (predictor steps) and Newton corrections are related to Newton steps, one may devise strategies to approximately compute the steps. We address in this section the asymptotic convergence order of such variants where both (predictors and correctors) steps are computed approximately.

A. Inexact SUMT

We now address inexact extrapolations and corrections. By solving approximately the equations defining the first order extrapolation $\hat{x}^1$, we get an extrapolate, denoted $\hat{x}$ with $\|\hat{x} - x(\rho^+, 0)\| \sim \rho^{a+1}$. If $a = 1$, we get as good a prediction as $\hat{x}^1$ while if $\hat{x}$ is computed cheaply, we insist to at least obtain an order $a + 1$ for some $a > 0$.

Assume that the Newton's direction is computed approximately such that $\nabla_x \Phi(\hat{x}, \rho^+) d_N + \Phi(\hat{x}, \rho^+) = R$ with $\|R\| \le \rho^{1+c} = \gamma$.

Lemma 3.1: Let $\hat{x}$ such that $\|\hat{x} - x(\rho^+, 0)\| \sim \rho^{a+1}$. Then, $d_N \sim \mathcal{O}(\gamma + \rho^{a+1})$.

1) Details for the log barrier: We first provide a proof of lemma 3.1

Proof: The primal-dual Newton's direction is written

$$\begin{pmatrix} \nabla_x L(\hat{x}, \hat{\lambda}) & \nabla \hat{g} \\ \hat{\Lambda} \nabla \hat{g}^t & \hat{G} \end{pmatrix} \begin{pmatrix} d_x \\ d_y \end{pmatrix} = \begin{pmatrix} -L(\hat{x}, \hat{\lambda}) + R \\ 0 = \hat{G}\hat{\lambda} - \rho^+ e \end{pmatrix}, \tag{13}$$

where $\hat{\lambda}_i = \frac{\rho^+}{\hat{g}_i}$. Define also $\lambda_i = \frac{\rho^+}{g_i}$, where $g_i = g_i(x(\rho^+, 0))$. By defining $\bar{d}_y = d_y - \lambda + \hat{\lambda}$, we may rewrite Eq.(13) as

$$\begin{pmatrix} \nabla_x L(\hat{x}, \hat{\lambda}) & \nabla \hat{g} \\ \hat{\Lambda} \nabla \hat{g}^t & \hat{G} \end{pmatrix} \begin{pmatrix} d_x \\ \bar{d}_y \end{pmatrix} = \begin{pmatrix} -L(\hat{x}, \lambda) + R \\ \hat{G}(\lambda - \hat{\lambda}) \end{pmatrix}, \tag{14}$$

We now observe that $L(\hat{x}, \lambda) = \mathcal{O}(\rho^{a+1})$ and $\hat{G}(\lambda - \hat{\lambda}) = \rho^+ G^{-1}(\hat{g} - g) = \mathcal{O}(\rho^{a+1})$ and $\|R\| \le \gamma$ which concludes the proof. ■

Now consider the effect of an inexact Newton correction.

Lemma 3.2: Let $\hat{x}$ such that $\|\hat{x} - x(\rho^+, 0)\| \sim \rho^{a+1}$. Then, $\Phi(\hat{x} + d_N, \rho^+) \sim \mathcal{O}\left(\frac{(\gamma + \rho^{(a+1)})^2}{\rho^{+2}}\right)$.

Proof: We write

$$\Phi(\hat{x} + d_N, \rho^+) = \Phi(\hat{x}, \rho^+) + \nabla_x \Phi(\hat{x}, \rho^+) d_N + \mathcal{O}\left(\frac{\|d_N\|^2}{(\rho^+)^2}\right), \tag{15}$$

noting the last denominator $(\rho^+)^2$ which comes from derivating twice $\frac{\rho^+}{g_i(\hat{x})}$. The first two terms are bounded by γ and using the bound on d_N from the lemma 3.1 we get $\frac{(\gamma + \rho^{(a+1)})^2}{\rho^{+2}}$. ■

Consider $\rho^+ = \rho^b$. If we want to be as cheap as possible while ensuring super-linear convergence ($b > 1$), we deduce that $b < \frac{2(1+c)}{3}$ while $c \le a$. Therefore, to get super-linear convergence, one has to pick $0.5 < c \le a$.

2) Details for quadratic loss: The proof of lemma 3.1 is very similar to the proof of the Lemma 3 in [2].

Now consider the effect of an inexact Newton correction.

Lemma 3.3: Let $\hat{x}$ such that $\|\hat{x} - x(\rho^+, 0)\| \sim \rho^{a+1}$. Then, $\Phi(\hat{x} + d_N, \rho^+) \sim \mathcal{O}\left(\frac{(\gamma + \rho^{(a+1)})^2}{\rho^+}\right)$.

Proof: We write

$$\Phi(\hat{x} + d_N, \rho^+) = \Phi(\hat{x}, \rho^+) + \nabla_x \Phi(\hat{x}, \rho^+) d_N + \mathcal{O}\left(\frac{\|d_N\|^2}{\rho^+}\right), \tag{16}$$

The first two terms are bounded by γ and using the bound on d_N from the lemma 3.1 we get $\frac{(\gamma + \rho^{(a+1)})^2}{\rho^+}$. ■

Consider $\rho^+ = \rho^b$. If we want to be as cheap as possible while ensuring super-linear convergence ($b > 1$), we deduce that $b < (1 + c)$ while $c \le a$. Therefore, to get super-linear convergence, one has to pick $0 < c \le a$.

IV. High order variants

Instead of solving approximately the predictor and-or corrector steps, we investigate here the effect of using higher order Taylor expressions of the central path. We reformulate slightly the equations to parametrize the path with the scalars ρ and τ. At the current point,

$$\Phi(x, \rho) = \tau \bar{\mathbf{r}} \tag{17}$$

for some residual vector $r = \tau \bar{\mathbf{r}}$ with $\bar{\mathbf{r}} = \frac{r}{\|r\|}$. Equation Eq.(17) induces a bi-parameter equation $x(\rho, \tau)$ and the solution searched for is $x^* = x(0, 0)$.

$$\begin{aligned} \hat{x}^1 &= x + \frac{\partial x}{\partial \rho}(\rho^+ - \rho) + \frac{\partial x}{\partial r}(-r) \\ \hat{x}^2 &= \frac{1}{2}\left(\frac{\partial^2 x}{\partial \rho^2}(\rho^+ - \rho)^2 + \frac{\partial^2 x}{\partial \rho \partial \tau}(\rho^+ - \rho)(-\tau) + \frac{\partial^2 x}{\partial \tau^2}(-\tau)^2\right) \\ \hat{x}^p &= \sum_{j=0}^{p} \binom{p}{j} \frac{\partial^p x}{\partial \rho^{p-j} \partial \tau^j} (\bar{\rho}^+ - \bar{\rho})^{p-j} (-\bar{\tau})^j \end{aligned}$$

We are now concerned with higher order extrapolations $\sum_{i=1}^{a} \hat{x}^i$ for $a > 1$. Postponing the actual computations of such a $\hat{x}^a$ for $a \ne 1$, we already may obtain the following.

Lemma 4.1: Let $\hat{x}$ such that $\|\hat{x} - x(\rho^+, 0)\| \sim \rho^{a+1}$ with $\frac{\rho^{a+1}}{\rho^+} < \infty$. Then, $\nabla \phi(\hat{x}, \rho^+) \sim \mathcal{O}(\frac{\rho^{a+1}}{\rho^+})$.

This result allows to claim that by using ($a > 1$)–order extrapolations, we get a $\frac{(a+1)}{2}$ order of convergence without even recourse to Newton corrections, and this both for the log barrier and the quadratic loss variants. Using a first order extrapolation is not enough, and requires a further Newton correction. Indeed, to reach the required approximation criterion,

$\nabla\phi(\hat{x},\rho^+)$ has to be lower than $\|r^+\| = \rho^+$, which implies that $\rho^{a+1} < {\rho^+}^2$.

Observe in particular that a second order extrapolation ($\hat{x}^1 + \hat{x}^2$) yields a predictor only algorithm achieving the limiting order $\frac{3}{2}$. This improves the exterior variant and even more so the interior variant since in this context, both interior and exterior variants share the same improved asymptotic behavior.

A. Proof of Lemma 4.1 for the log–barrier

Proof: Denote $x = x(\rho^+, 0)$; then $\Phi(x, \rho^+) = 0$ and write:

$$\begin{aligned}
\Phi(\hat{x},\rho^+) &= \Phi(\hat{x},\rho^+) - \Phi(x,\rho^+) \\
&= \nabla\hat{f} - \nabla f + \sum \frac{\rho^+}{\hat{g}_i}\nabla\hat{g}_i - \sum\frac{\rho^+}{g_i}\nabla g_i \\
&= \mathcal{O}(\rho^{a+1}) + \sum\frac{\rho^+(g_i-\hat{g}_i)}{\hat{g}_i g_i}\nabla\hat{g}_i \\
&\quad + \sum\frac{\rho^+}{g_i}(\nabla\hat{g}_i - \nabla g_i) \\
&= \mathcal{O}(\rho^{a+1}) + \sum\frac{\mathcal{O}(\rho^{a+1})\rho^+}{\hat{g}_i g_i}\nabla\hat{g}_i \\
&\quad + \sum\lambda_i\mathcal{O}(\rho^{a+1}).
\end{aligned}$$

We have $g_i \sim \Theta(\rho^+)$ and $\hat{g}_i = g_i + \mathcal{O}(\rho^{a+1})$, so that $\hat{g}_i \sim \Theta(\rho^+)$ since $\frac{\rho^{a+1}}{\rho^+}$ is bounded. ■

B. Proof of Lemma 4.1 for the quadratic loss

Proof: Denote $x = x(\rho^+, 0)$; then $\Phi(x, \rho^+) = 0$ and write:

$$\begin{aligned}
\Phi(\hat{x},\rho^+) &= \Phi(\hat{x},\rho^+) - \Phi(x,\rho^+) \\
&= \nabla\hat{f} - \nabla f + \sum \frac{\hat{g}_i}{\rho^+}\nabla\hat{g}_i - \sum\frac{g_i}{\rho^+}\nabla g_i \\
&= \mathcal{O}(\rho^{a+1}) + \frac{(g_i-\hat{g}_i)}{\rho^+}\nabla\hat{g}_i \\
&\quad + \sum\frac{g_i}{\rho^+}(\nabla\hat{g}_i - \nabla g_i) \\
&= \mathcal{O}(\rho^{a+1}) + \frac{\mathcal{O}(\rho^{a+1})}{\rho^+}\nabla\hat{g}_i \\
&\quad + \sum\lambda_i\mathcal{O}(\rho^{a+1}).
\end{aligned}$$

We have $g_i \sim \Theta(\rho^+)$ and $\hat{g}_i = g_i + \mathcal{O}(\rho^{a+1})$, so that $\hat{g}_i \sim \Theta(\rho^+)$ since $\frac{\rho^{a+1}}{\rho^+}$ is bounded. ■

V. Computing extrapolations

The actual extrapolation computations for the quadratic loss function is presented in details in [3]. To avoid derivatives of the $\frac{1}{\rho}$ factor involved in the penalty term, we resorted to primal-dual equations in [3]. We develop in this section the details for the high order derivatives for the log barrier.

We rewrite equation Eq.(10) in a simplified notation:

$$\Phi(x,\rho) = c - \rho A^t G^{-1} e. \tag{18}$$

For linear programs, c and A are constant while otherwise, $c = \nabla f(x)$ and $A = \nabla g(x)$. $G = \mathrm{diag}(g_i(x))$, and for linear programs, $g(x) = Ax - b$. We note for the sequel

$$\begin{aligned}
\nabla_x\Phi(x,\rho) &= \rho A^t G^{-2} A + \nabla^2_{xx} l(x, \rho G(x)^{-1}e) && (19)\\
\nabla_\rho\Phi(x,\rho) &= -A^t G^{-1} e && (20)
\end{aligned}$$

and remark that the Lagrangian term vanishes for linear programs.

The implicit function theorem yields

$$\begin{aligned}
\nabla_x\Phi(x,\rho)\dot{x}_\rho + \nabla_\rho\Phi(x,\rho) &= 0 \\
\nabla_x\Phi(x,\rho)\dot{x}_\tau - \bar{\mathbf{r}} &= 0.
\end{aligned} \tag{21}$$

Thus, the combined extrapolation step reduces to

$$\begin{aligned}
&\nabla_x\Phi(x,\rho)(\dot{x}_\rho(\rho^+ - \rho)) + \dot{x}_\tau(-\tau) \\
&\quad + \nabla_\rho\Phi(x,\rho)(\rho^+ - \rho) + \tau\bar{\mathbf{r}} = 0,
\end{aligned} \tag{22}$$

which, for this first order candidate, simplifies to $\nabla_x\Phi(x,\rho)\hat{x}^1 + \Phi(x,\rho^+) = 0$.

In order to go further to the expressions of higher order extrapolates, we first note the following for the log barrier case:

$$\begin{aligned}
\nabla^2_{x\rho}\Phi(x,\rho) = \nabla^2_{\rho x}\Phi(x,\rho) &= A^t G^{-2} A + \nabla^2_{xx} l(x, G(x)^{-1}e) \\
\nabla^2_{\rho\rho}\Phi &\equiv 0 \\
\nabla^2_{\tau\cdot}\Phi = \nabla^2_{\cdot\tau}\Phi &\equiv 0
\end{aligned}$$

In a nutshell, any derivative of Φ with respect to τ vanishes since Φ does not involve τ, and any high order derivative of Φ with respect to ρ more than once also vanishes since Φ is linear in ρ.

Now, still using the implicit function theorem, this time to equations Eq.(21), we get the following, in which we use Φ without arguments as a shorthand notation for $\Phi(x,\rho)$:

$$\begin{aligned}
\nabla^2_{xx}\Phi\dot{x}_\rho\dot{x}_\rho + \left(\nabla^2_{x\rho}\Phi + \nabla^2_{\rho x}\Phi\right)\dot{x}_\rho + \nabla_x\Phi\ddot{x}_{\rho\rho} &= 0 \\
\nabla^2_{xx}\Phi\dot{x}_\tau\dot{x}_\rho + \nabla^2_{\rho x}\Phi\dot{x}_\tau + \nabla_x\Phi\ddot{x}_{\tau\rho} &= 0 \\
\nabla^2_{xx}\Phi\dot{x}_\rho\dot{x}_\tau + \nabla^2_{x\rho}\Phi\dot{x}_\tau + \nabla_x\Phi\ddot{x}_{\rho\tau} &= 0 \\
\nabla^2_{xx}\Phi\dot{x}_\tau\dot{x}_\tau + \nabla_x\Phi\ddot{x}_{\tau\tau} &= 0
\end{aligned} \tag{23}$$

Observe that the four relations above all imply a linear system defined by the same matrix $\nabla_x\Phi(x,\rho)$ and the following four right hand sides, conveniently expressed using $\bar{x}_\tau$ which denotes a constant vector of value $\dot{x}_\tau$, and similarly $\bar{x}_\rho$ is a constant vector of value $\dot{x}_\rho$:

$$\begin{aligned}
\nabla_\rho\left(\nabla_x\Phi\bar{x}_\rho + \nabla_\rho\Phi\right) &= \nabla^2_{xx}\Phi\dot{x}_\rho\bar{x}_\rho + \nabla^2_{x\rho}\Phi\bar{x}_\rho + \nabla^2_{\rho x}\Phi\dot{x}_\rho && (24)\\
\nabla_\tau\left(\nabla_x\Phi\bar{x}_\rho + \nabla_\rho\Phi\right) &= \nabla^2_{xx}\Phi\dot{x}_\tau\bar{x}_\rho + \nabla^2_{\rho x}\Phi\dot{x}_\tau && (25)\\
\nabla_\rho\left(\nabla_x\Phi\bar{x}_\tau - \bar{\mathbf{r}}\right) &= \nabla^2_{xx}\Phi\dot{x}_\rho\bar{x}_\tau + \nabla^2_{x\rho}\Phi\bar{x}_\tau && (26)\\
\nabla_\tau\left(\nabla_x\Phi\bar{x}_\tau - \bar{\mathbf{r}}\right) &= \nabla^2_{xx}\Phi\dot{x}_\tau\bar{x}_\tau && (27)
\end{aligned}$$

Hereafter, we use the "bar" $\bar{\rho}$ and $\bar{\tau}$ to represent actual extrapolation steps values, as opposed to variables within the equations. Now, $\hat{x}^2 = \ddot{x}_{\rho\rho}(\bar{\rho}^+ - \bar{\rho})^2 + 2\ddot{x}_{\tau\rho}(\bar{\rho}^+ - \bar{\rho})(-\bar{\tau}) + \ddot{x}_{\tau\tau}(\bar{\tau})^2$ so that the right hand sides involving the second derivatives may be combined into $(\bar{\rho}^+ - \bar{\rho})((\bar{\rho}^+ - \bar{\rho})Eq.(24) - \bar{\tau}Eq.(26))$

and $-\bar{\tau}((\bar{\rho}^+ - \bar{\rho})Eq.(25) - \bar{\tau}Eq.(27))$ and, also using the notation that $\hat{\bar{x}}^1$ is a constant vector of value $\hat{x}^1$, is expressed:

$$\nabla_\rho \left(\nabla_x \Phi(x,\rho)\hat{\bar{x}}^1 + \nabla_\rho \Phi(x,\rho)(\bar{\rho}^+ - \bar{\rho}) - \bar{\mathbf{r}}\bar{\tau}\right)(\bar{\rho}^+ - \bar{\rho}) + \nabla_\tau \left(\nabla_x \Phi(x,\rho)\hat{\bar{x}}^1 + \nabla_\rho \Phi(x,\rho)(\bar{\rho}^+ - \bar{\rho}) - \mathbf{r}\bar{\tau}\right)(-\tau) \tag{28}$$

To establish a recurrence relation to compute the $\hat{x}^p$, it is convenient to define a family of functions

$$\Phi^0(x,\rho,\tau) = \Phi(x,\rho) - \tau\bar{\mathbf{r}} \tag{29}$$

$$\Phi^p(x,\rho,\tau) = (\bar{\rho}^+ - \bar{\rho})\nabla_\rho \Phi^{p-1}(x,\rho,\tau) - \bar{\tau}\nabla_\tau \Phi^{p-1}(x,\rho,\tau) \tag{30}$$

Theorem 5.1:

$$\hat{x}^p = \sum_{j=0}^{p} \binom{p}{j} \frac{\partial^p x}{\partial \rho^{p-j} \partial \tau^j} (\bar{\rho}^+ - \bar{\rho})^{p-j} (-\bar{\tau})^j$$

satisfies $\nabla_x \Phi(x,\rho)\hat{x}^p + \Phi^p(x,\rho,\tau) = 0$.

Proof: The inductive proof has its base verified by the relation Eq.(28). The induction step will use the relation:

$$\hat{x}^{p+1} = \frac{\partial \hat{x}^p}{\partial \rho}(\bar{\rho}^+ - \bar{\rho}) + \frac{\partial \hat{x}^p}{\partial \tau}(-\bar{\tau})$$

The equations Φ^p includes a term $\nabla_x \Phi(x,\rho)\hat{x}^p$ defining the linear system, the remaining of Φ^p corresponding to the right hand side of the linear equation. ■

The recurrence Φ^p may be explicitly written as

$$\nabla_x \Phi(x,\rho)\hat{x}^p + \hat{\Phi}^p$$

where $\hat{\Phi}^p$ involves terms of the form $\nabla^j_{x^{j_x}\rho^{j_\rho}} \Phi(x,\rho) v_1^{i_1} v_2^{i_2} \dots v_l^{i_l}$ with $\sum_{k=1}^{l} i_k = j_x$, $j_x + j_\rho = j$ and $1 < j \le p$. Moreover, each v_k is composed of partial derivatives of x with respect to ρ and/or τ up to order $j_x - 1$. As it happens, the recurrence may be written using only the $\hat{x}^p$ without explicit reference to the (mixed) partials derivatives of x wrt ρ or τ:

$$\Phi^0(x,\rho,\tau) = \Phi(x,\rho) + \tau\bar{\mathbf{r}} \tag{31}$$

$$\Phi^1(x,\rho,\tau) = \nabla_x \Phi(x,\rho)\hat{x}^1 + (\bar{\rho}^+ - \bar{\rho})\nabla_\rho \Phi(x,\rho) + \bar{\tau}\bar{\mathbf{r}} \tag{32}$$

$$\begin{aligned}\Phi^2(x,\rho,\tau) = {} & \nabla_x \Phi(x,\rho)\hat{x}^2 + 2(\bar{\rho}^+ - \bar{\rho})\nabla^2_{x\rho}\Phi(x,\rho)\hat{x}^1 \\ & + \nabla^2_{xx}\Phi(x,\rho)\hat{x}^1\hat{x}^1\end{aligned} \tag{33}$$

$$\begin{aligned}\Phi^3(x,\rho,\tau) = {} & \nabla_x \Phi(x,\rho)\hat{x}^3 + 3(\bar{\rho}^+ - \bar{\rho})\nabla^2_{x\rho}\Phi(x,\rho)\hat{x}^2 \\ & + 3\nabla^2_{xx}\Phi(x,\rho)\hat{x}^1\hat{x}^2 \\ & + 3(\bar{\rho}^+ - \bar{\rho})\nabla^3_{xx\rho}\Phi(x,\rho)\hat{x}^1\hat{x}^1 \\ & + \nabla^3_{xxx}\Phi(x,\rho)\hat{x}^1\hat{x}^1\hat{x}^1\end{aligned} \tag{34}$$

$$\begin{aligned}\Phi^4(x,\rho,\tau) = {} & \nabla_x \Phi(x,\rho)\hat{x}^4 + 4(\bar{\rho}^+ - \bar{\rho})\nabla^2_{x\rho}\Phi(x,\rho)\hat{x}^3 \\ & + 3\nabla^2_{xx}\Phi(x,\rho)\hat{x}^2\hat{x}^2 \\ & + 4\nabla^2_{xx}\Phi(x,\rho)\hat{x}^1\hat{x}^3 \\ & + 12(\bar{\rho}^+ - \bar{\rho})\nabla^3_{xx\rho}\Phi(x,\rho)\hat{x}^1\hat{x}^2 \\ & + 6\nabla^3_{xxx}\Phi(x,\rho)\hat{x}^2\hat{x}^1\hat{x}^1 \\ & + 4(\bar{\rho}^+ - \bar{\rho})\nabla^4_{xxx\rho}\Phi(x,\rho)\hat{x}^1\hat{x}^1\hat{x}^1 \\ & + \nabla^4_{xxxx}\Phi(x,\rho)(\hat{x}^1)^4\end{aligned} \tag{35}$$

1) Implementation for linear programming: By introducing the notation $v^1 = A\hat{x}^1$, and $V = \mathrm{diag}(v)$, we may express the high order terms using the following lemma.

Lemma 5.2:

$$\nabla_x (v^t V_p G^{-p} u) = -pA^t V V_p G^{-(p+1)} u \tag{36}$$

This allows to write equation Eq.(33) as

$$\rho A^t G^{-2} A\hat{x}^2 - 2\rho A^t V^1 G^{-3} v^1 + 2(\bar{\rho}^+ - \bar{\rho})A^t G^{-2} v^1 = 0 \tag{37}$$

Similarly, equation Eq.(34) leads to the following expression:

$$\begin{aligned} & -6\rho A^t V^1 G^{-3} v^2 - 6(\bar{\rho}^+ - \bar{\rho})A^t V^1 G^{-3} v^1 \\ & + 3(\bar{\rho}^+ - \bar{\rho})A^t G^{-2} v^2 + 6\rho A^t (V^1)^2 G^{-4} v^1\end{aligned} \tag{38}$$

As we may observe, each term involves a single matrix–vector computation in addition to several $\mathcal{O}(n)$ diagonal matrices and vector operations, overall yielding cheap right hand sides to compute higher order derivatives. This was to be expected.

2) General implementation using automatic differentiation (AD): Using AD tools, we may evaluate higher order derivative cheaply too. Assuming full dense Hessian's—constraint jacobians, the linear system requires $\mathcal{O}(n^3)$ arithmetic operations to factorizes, and further on, back–front substitutions together with right hand side computations reduce to $\mathcal{O}(n^2)$ arithmetic operations. As it happens, we may get the high order right hand sides required for the Taylor coefficients in $\mathcal{O}(n^2)$ complexity, leaving the main burden to obtain and factorize the Jacobian Matrix.

VI. Shamanskii inspired extrapolations

In the context of unconstrained optimization using Newton's method, reusing the Hessian matrix is closely related to Shamanskii's method, sometimes refered to as composite Newton method. Shamanskii's consists in reusing the Hessian two or more times; this is interesting since factorizing the Hessian has a much superior computational cost than using the factorization to solve a linear system. From an asymptotic point of view, then, high order extrapolations (reminiscent of Chebychev method) or Shamanskii method share the same improvement with respect to the convergence order. The simplicity of Shamanskii's approach is appealing.

We will provide a Shamanskii approximation to the second order extrapolation. This yields an approximate second order predictor algorithm reaching the limiting convergence order $\frac{3}{2}$.

We analyze the following scheme.

$$\begin{aligned}\nabla_x \Phi(x,\rho)\check{x}^1 + \Phi(x,\rho^+) &= 0 \\ \nabla_x \Phi(x,\rho)\check{x}^2 + \Phi(x + \check{x}^1,\rho^+) &= 0\end{aligned}$$

Thus, we reuse $\nabla_x \Phi(x,\rho)$ and only change the right hand sides. We recognize that $\check{x}^1 = \hat{x}^1$ previously defined.

Proposition 6.1: The second order $\check{x}^2$ is an $\mathcal{O}(\rho^3)$ approximation to the second order extrapolation $\hat{x}^1 + \hat{x}^2$.

Proof: We express the right hand side using a Taylor expansion. All the terms in the right hand side are functions evaluated at x and ρ, and thus the various $\Phi(x,\rho)$ will be shorthanded to Φ.

$$\begin{aligned} \Phi(x+\check{x}^1,\rho^+) &= \Phi \\ &+ \nabla_x \Phi \check{x}^1 + \nabla_\rho \Phi(\rho^+ - \rho) && (39)\\ &+ \nabla^2_{xx}\Phi \check{x}^1 \check{x}^1 && (40)\\ &+ 2\nabla^2_{x\rho}\Phi \check{x}^1(\rho^+ - \rho) && (41)\\ &+ \nabla^2_{\rho\rho}\Phi(\rho^+ - \rho)^2 && (42)\\ &+ \mathcal{O}(\max(\tau,\rho)^3) \end{aligned}$$

Now, we already have seen that $\nabla^2_{\rho\rho}\Phi = 0$, and by the definition of $\check{x}^1$, Eq.(39) vanishes, which yields that $\hat{x}^2$ Eq.(33) and $\check{x}^2$ Eq.(40) and Eq.(41) differ by $\mathcal{O}(\rho^3)$. ■

The second order Shamanskii direction is thus a suitable approximation of the second order extrapolation. The conclusion of the lemma 4.1 then still holds.

Corollary 6.2: $\nabla\phi(\check{x}^2,\rho^+) \sim \mathcal{O}(\frac{\rho^3}{\rho^+})$

The process may be continued,

$$\nabla_x \Phi(x,\rho)\check{x}^3 + \Phi(x+\check{x}^1+\check{x}^2,\rho^+) = 0,$$

and in general,

$$\nabla_x \Phi(x,\rho)\check{x}^{p+1} + \Phi(x+\sum_{i=1}^{p}\check{x}^i,\rho^+) = 0.$$

We conjecture that the $\check{x}^p$ may be used and preserve the good properties of the $\sum_{i=1}^{p}\hat{x}^i$.

VII. Numerical illustration

We now provide a simple numerical example to exhibit the benefits of using a Shamanskii like extrapolation.

We consider the simple example

$$\begin{aligned} \min_{x\in\mathbb{R}^6} \quad f(x) &= \sum_{i=1}^{6} i x^i \\ s.t. \quad g_1(x) &= (x_1+x_3+x_5)^2 - 1 = 0 \\ g_2(x) &= (x_2+x_3+x_4)^2 - 1 = 0 \\ g_3(x) &= x_1 x_6 - 1 = 0 \end{aligned}$$

We use the quadratic penalty function. Therefore, we hope two-steps superlinear convergence almost quadratic using a first order extrapolation and almost $\frac{3}{2}$ convergence order using a Shamanskii variant. The two steps in the first order variant require factorization and solution of two distinct linear systems while the Shamanskii variant uses a single factorization to solve two related linear systems.

We compare a sub-$\frac{3}{2}$ sequence ρ_k using both a first order extrapolation and a Shamanskii–2 extrapolation. We may observe in table I that the extrapolation does not require any Newton correction for the last four two-steps extrapolations. In the tables, the first 7 iterations are identical and thus are omitted. We observe that the first order variant requires a few more iterations to reach our (tight) tolerance.

The first order variant could be improved by considering a sub-quadratic sequence for ρ_k, and we exhibit the results in table III. The first 13 iterations are identical with those from table II, and we confirm that overall, this variant betters the slower sequence ρ_k, but the Shamanskii variant is still the most efficient. The limiting order when considering two factorizations is quadratic for the linear extrapolation and $\frac{9}{4}$ for the Shamanskii version, which is coherent with our example.

As a final remark regarding this simple illustration, the quadratic penalty using Shamanskii strategy benefits much less than log-barrier algorithms. Nevertheless, our example suggests that it (Shamanskii) may improve upon the plain first order extrapolation.

TABLE I
SECOND ORDER SUB $\frac{3}{2}$ VARIANT

ρ	Iter	$\|\nabla p(x,\rho)\|$	$\|g(x)\|$	∇L
Ex two	7	2.2e+00	5.8e-02	
1.0e-02				
Nwt	8	1.2e-02	5.5e-02	
Nwt	9	6.5e-05	5.5e-02	
Ex two	10	2.7e-02	5.6e-03	
1.0e-03				
Nwt	11	1.1e-06	5.6e-03	
Ex two	12	2.7e-04	5.6e-04	
1.0e-04				
Nwt	13	1.2e-09	5.6e-04	
Ex two	14	2.7e-06	5.6e-05	
1.0e-05				
Ex two	15	2.8e-08	5.6e-06	
1.0e-06				
Ex two	16	9.8e-08	5.6e-08	5.5e-12
1.0e-08				
Ex two	17	6.9e-05	5.6e-11	1.2e-15
1.0e-11				
Ex two	18	1.2e-01	5.6e-14	1.2e-15

VIII. Conclusion and future work

In this paper, we summarized known results about the asymptotic behavior of SUMT algorithms in non-linear optimization. We considered both interior and exterior penalty variants. Overall, exterior variants enjoy better asymptotic properties.

As can be seen from the table, interior variants suffer from poorer asymptotic convergence order. In particular, for the inexact predictor-corrector strategy, one has to impose $c > 0.5$ i.e. the residual of the extrapolation and the Newton correction has to be reduced to an order at least 1.5 merely to provide an overall two-steps superlinear behavior. The exterior variant is somewhat more forgiving in this context.

High order predictors make interior and exterior approaches competitive. Actually, the use of higher order predictors allow to bypass any Newton corrector step asymptotically, and for both the interior and exterior variant, allow to reach the same order of convergence limit, namely $\frac{k+1}{2}$ for order $k > 1$ extrapolates.

TABLE II
FIRST ORDER SUB $\frac{3}{2}$ VARIANT

ρ	Iter	$\|\nabla p(x,\rho)\|$	$\|g(x)\|$	∇L
Ex one 1.0e-02	7	1.3e+01	1.1e-01	
Nwt	8	4.7e-01	5.3e-02	
Nwt	9	4.4e-04	5.5e-02	
Ex one 1.0e-03	10	1.7e+00	5.8e-03	
Nwt	11	7.4e-04	5.6e-03	
Ex one 1.0e-04	12	1.7e-01	5.6e-04	
Nwt	13	7.4e-07	5.6e-04	
Ex one 1.0e-05	14	1.8e-02	5.6e-05	
Nwt	15	7.0e-10	5.6e-05	
Ex one 1.0e-06	16	1.8e-03	5.6e-06	
Nwt	17	1.6e-09	5.6e-06	
Ex one 1.0e-08	18	2.2e-03	5.6e-08	7.2e-12
Ex one 1.0e-11	19	2.2e+00	5.4e-11	1.6e-12
Ex one 1.0e-14	20	2.2e+03	7.4e-12	1.2e-12
Nwt	21	3.1e-02	5.6e-14	1.5e-11
Nwt	22	3.1e-02	5.6e-14	3.4e-15

TABLE III
FIRST ORDER SUB QUADRATIC VARIANT

ρ	Iter	$\|\nabla p(x,\rho)\|$	$\|g(x)\|$	∇L
Ex one 1.0e-06	14	2.2e-01	5.6e-06	
Nwt	15	1.0e-08	5.6e-06	
Ex one 1.0e-09	16	2.2e-02	5.6e-09	7.3e-12
Ex one 1.0e-14	17	2.2e+03	7.5e-12	1.7e-12
Nwt	18	1.8e-01	5.6e-14	1.7e-11
Nwt	19	3.1e-02	5.6e-14	4.8e-15

It should be recalled that order–k predictor incur a computational cost of $\mathcal{O}(n^3)$ arithmetic operation to factorize the jacobian matrix plus k times $\mathcal{O}(n^2)$ to obtain the high order terms while the corrector steps involve the solution of a linear system, again $\mathcal{O}(n^3)$. Therefore, from a complexity per iteration point of view, high order predictors are far preferable to their first order predictor-corrector counterpart: they require only one $\mathcal{O}(n^3)$ factorization and k $\mathcal{O}(n^2)$ substitutions while the first order approach requires two $\mathcal{O}(n^3)$ factorization and two $\mathcal{O}(n^2)$ substitutions.

The results presented suggest that the use of the Shamanskii approximation to the higher order trajectory derivatives is a simple solution to reach good asymptotic convergence properties, equivalently good for interior and exterior variants of SUMT. This contrasts with the usage of a simple extrapolation, or approximate computations of the predictor and corrector steps.

The analysis may be combined into a mixed penalty approach to treat programs involving both equality constraints and inequalities. This new strategy outperforms (from an asymptotic analysis point of view) previous studies using mixed interior and exterior penalties. The analysis also may be applied to the exponential penalty as well as other variations.

Future works will involve implementation and comparisons with primal-dual methods. Primal-dual interior point methods have very good convergence properties, but require skill to ensure global convergence while exhibiting good asymptotic behavior.

TABLE IV
LIMITING CONVERGENCE ORDERS FOR VARIANTS DISCUSSED IN THE PAPER

variant	interior	exterior
corrector	linear	linear
predictor-corrector[1]	2–steps $\frac{4}{3}$	2–steps quadratic
order-$k \geq 2$ pred	$\frac{k+1}{2}$	$\frac{k+1}{2}$
inexact pred-corr	2–steps–$\frac{2}{3}(1+c)$	2–steps $1+c$

REFERENCES

[1] Jean-Pierre Dussault. Numerical stability and efficiency of penalty algorithms. *S.I.A.M. Journal on Numerical Analysis*, 32(1):296–317, February 1995.

[2] Jean-Pierre Dussault. Augmented penalty algorithms. *IMA Journal on Numerical Analysis*, 18:355–372, 1998.

[3] Jean-Pierre Dussault. High order Newton-penalty algorithms. *Journal of computational and applied mathematics*, 182(1):117–133, oct 2005.

[4] Jean-Pierre Dussault and Abdelatif Elafia. On the convergence rate of the logarithmic barrier algorithm. *Computational Optimization and Applications*, 19(1):31–54, apr 2001.

[5] Antony V. Fiacco and Garth P. McCormick. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. S.I.A.M., 1990.

[6] Sanjay Mehrotra. Asymptotic convergence in a generalized predictor-corrector method. *Math. Program.*, 74(1):11–28, 1996.