Convergence Analysis of a Trust-Region Method under Noisy Settings

Liyuan Cao

Lehigh University ⇒ Peking University

October 23, 2021

Collaborators

Albert S. Berahas
U of Michigan

Katya Scheinberg
Cornell University

Trust Region Method

$$\min_{x \in \mathbb{R}^n} \phi(x)$$

ϕ is smooth but we only have access to

$$\begin{cases} f(x) = \phi(x) + e(x) \\ g_k \\ H_k \end{cases} \quad \text{instead of} \quad \begin{cases} \phi(x) \\ \nabla \phi(x) \\ \nabla^2 \phi(x) \end{cases}$$

Algorithm: Modified Trust Region Algorithm

Inputs: Starting point x_0, initial trust region radius δ_0, and hyperparameters $\eta_1 > 0, \eta_2 > 0$, $\tau \in (0,1)$ for controlling the trust region radius.

for $k = 0, 1, 2, \ldots$ **do**

Build a quadratic model $m_k(x_k + s) = \phi(x_k) + \langle g_k, s\rangle + 0.5\,\langle H_k s, s\rangle$

Compute s_k by approximately minimizing m_k in $B(x_k, \delta_k)$ so that it satisfies the *Cauchy decrease condition.*

Compute

$$\rho_k = \frac{f(x_k) - f(x_k + s_k) + r}{m_k(x_k) - m_k(x_k + s_k)}$$

if $\rho_k \geq \eta_1$ **then**

Set $x_{k+1} = x_k + s_k$ and

$$\delta_{k+1} = \begin{cases} \tau^{-1}\delta_k & \text{if } \|g_k\| \geq \eta_2\delta_k \\ \tau\delta_k, & \text{if } \|g_k\| < \eta_2\delta_k \end{cases}$$

else

Set $x_{k+1} = x_k$ and $\delta_{k+1} = \tau\delta_k$

Assumptions

$\|\nabla\phi(x) - \nabla\phi(y)\| \leq L\|x - y\|$ for all $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$
$\phi(x) \geq \hat{\phi}$ for all $x \in \mathbb{R}^n$.

random variables:	X_k	X_k^+	$\mathcal{E}_k$	$\mathcal{E}_k^+$	
realizations:	x_k	$x_k + s_k$	$\lvert e(x_k)\rvert$	$\lvert e(x_k + s_k)\rvert$	
random variables:	G_k	$\mathcal{H}_k$	M_k	Δ_k	ρ_k
realizations:	g_k	H_k	m_k	δ_k	ρ_k

We assume for all $k = 0, 1, \ldots$:

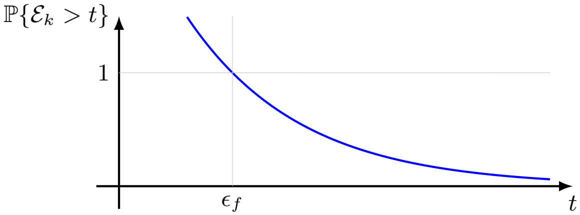
- $\|\mathcal{H}_k\| \leq \kappa_{\text{bhm}}$ for some constant κ_{bhm} (bound on hessian of model) deterministically;
- $\mathbb{P}\{\|G_k - \nabla\phi(X_k)\| \leq \kappa_{\text{eg}}\Delta_k \mid \mathcal{F}_{k-1}\} \geq p_1$ for some constant κ_{eg} (error of gradient) and $p_1 > 0.5$, where $\mathcal{F}_{k-1}$ is the sigma-algebra generated by all the random events before iteration k;
- $\left.\begin{array}{l}\mathbb{P}\{\mathcal{E}_k > t\} \\ \mathbb{P}\{\mathcal{E}_k^+ > t\}\end{array}\right\} \leq \exp(a(\epsilon_f - t))$ and $\{(\mathcal{E}_k, \mathcal{E}_k^+)\}$ independent.

Assumptions

$\mathbb{P}\{\|G_k - \nabla\phi(X_k)\| \le \kappa_{\text{eg}}\Delta_k \mid \mathcal{F}_{k-1}\} \ge p_1$

- Sample size needs to increase when Δ_k is small.

$\left.\begin{array}{l}\mathbb{P}\{\mathcal{E}_k > t\} \\ \mathbb{P}\{\mathcal{E}_k^+ > t\}\end{array}\right\} \le \exp(a(\epsilon_f - t))$ and $\{(\mathcal{E}_k, \mathcal{E}_k^+)\}$ independent.

- 

- $\mathcal{E}_k$ can be dependent $\mathcal{E}_k^+$, e.g. evaluated with the same sample set in an ERM setting.

Goal

While first-order convergence analyses for nonconvex functions typically work towards a result of the form

$$\min\{\|\nabla\phi(x_k)\| : \ 0 \le k \le T-1\} \le \begin{array}{l}\text{a function of } T \text{ that} \\ \text{converges to 0 as } T \text{ increases,}\end{array}$$

we could not have that result due to stochasticity. Instead, the main goal is to derive a probabilistic result of the form

$$\mathbb{P}\left\{\min\{\|\nabla\phi(X_k)\| : \ 0 \le k \le T-1\} < \epsilon\right\}$$
$$\ge \text{a function of } T \text{ the converges to 1 as } T \text{ increase}$$

for any sufficiently large ϵ.

Lemmas for Individual Iterations

Lemma (**sufficient condition for successful step**)

If $\|g_k - \nabla\phi(x_k)\| \le \kappa_{\text{eg}}\delta_k$ *and* $e(x_k) - e(x_k + s_k) + r \ge 0$, *and*

$$\delta_k \le \min\left\{ \frac{1}{\kappa_{eg} + \frac{L+\kappa_{\text{bhm}}+2\kappa_{\text{eg}}}{\kappa_{fcd}(1-\eta_1)}}, \frac{1}{\kappa_{eg}+\eta_2} \right\} \|\nabla\phi(x_k)\| := C_1\|\nabla\phi(x_k)\|,$$

then $\rho_k \ge \eta_1$ *and* $\|g_k\| \ge \eta_2\delta_k$.

Lemma (**progress made in each iteration**)

Let $h(\delta) = \frac{1}{2}\eta_1\eta_2\kappa_{fcd}\min\left\{\frac{\eta_2}{\kappa_{bhm}}, 1\right\}\delta^2 := C_2\delta^2$. *Then we have*

$$\phi(x_k) - \phi(x_{k+1}) \ge \begin{cases} h(\delta_k) - e(x_k) + e(x_k+s_k) - r & \text{if } \rho_k \ge \eta_1 \text{ and } \|g_k\| \ge \eta_2\delta_k \\ -e(x_k) + e(x_k+s_k) - r & \text{if } \rho_k \ge \eta_1 \\ 0 & \text{if } \rho_k < \eta_1. \end{cases}$$

Classification of Iterations

Lemma (**sufficient condition for successful step**)

If $\|g_k - \nabla\phi(x_k)\| \leq \kappa_{\text{eg}}\delta_k$ *and* $e(x_k) - e(x_k + s_k) + r \geq 0$, *and*

$$\delta_k \leq \min\left\{\frac{1}{\kappa_{eg} + \frac{L+\kappa_{\text{bhm}}+2\kappa_{\text{eg}}}{\kappa_{fcd}(1-\eta_1)}}, \frac{1}{\kappa_{eg}+\eta_2}\right\}\|\nabla\phi(x_k)\| := C_1\|\nabla\phi(x_k)\|,$$

then $\rho_k \geq \eta_1$ *and* $\|g_k\| \geq \eta_2\delta_k$.

Define

$$I_k = \mathbb{1}\{\|G_k - \nabla\phi(X_k)\| \leq \kappa_{\text{eg}}\Delta_k\}$$

$$\mathbb{P}\{I_k = 1 \mid \mathcal{F}_{k-1}\} \geq p_1 \text{ assumed}$$

$$J_k = \mathbb{1}\{\mathcal{E}_k + \mathcal{E}_k^+ \leq r\}$$

$$\Lambda_k = \mathbb{1}\{\Delta_k > \bar{\Delta}\}$$

$$\Lambda'_k = \mathbb{1}\{\Delta_k > \tau\bar{\Delta}\}$$

where $\bar{\Delta} = C_1 \min\{\|\nabla\phi(X_k)\| : \; k = 1, 2, \ldots, T-1\}$.

Classification of Iterations

Define

$$\Theta_k = \mathbb{1}\{\rho_k \geq \eta_1 \text{ and } \|G_k\| \geq \eta_2 \Delta_k\}$$
$$\Theta'_k = \mathbb{1}\{\rho_k \geq \eta_1\}$$

Lemma (**progress made in each iteration**)

Let $h(\delta) = \frac{1}{2}\eta_1\eta_2\kappa_{fcd}\min\left\{\frac{\eta_2}{\kappa_{bhm}}, 1\right\}\delta^2 := C_2\delta^2$. *Then we have*

$$\phi(x_k) - \phi(x_{k+1}) \geq$$
$$\begin{cases} h(\delta_k) - e(x_k) + e(x_k + s_k) - r & \text{if } \rho_k \geq \eta_1 \text{ and } \|g_k\| \geq \eta_2\delta_k \\ -e(x_k) + e(x_k + s_k) - r & \text{if } \rho_k \geq \eta_1 \\ 0 & \text{if } \rho_k < \eta_1. \end{cases}$$

Lemmas

Lemma (sufficient condition for successful step)

If $I_k J_k = 1$ *and* $\Lambda_k = 0$ *then* $\Theta_k = 1$.

Lemma (progress made in each iteration)

Let $h(\delta) = \frac{1}{2}\eta_1\eta_2\kappa_{fcd}\min\left\{\frac{\eta_2}{\kappa_{bhm}}, 1\right\}\delta^2 := C_2\delta^2$. *Then we have*

$$\phi(X_k) - \phi(X_{k+1}) \geq \begin{cases} h(\Delta_k) - \mathcal{E}_k - \mathcal{E}_k^+ - r & \textit{if } \Theta_k = 1 \\ -\mathcal{E}_k - \mathcal{E}_k^+ - r & \textit{if } \Theta_k' = 1 \\ 0 & \textit{if } \Theta_k' = 0. \end{cases}$$

Lemmas

Lemma (sufficient condition for successful step)

If $I_k J_k = 1$ *and* $\Lambda_k = 0$ *then* $\Theta_k = 1$.

Lemma (progress made in each iteration)

Let $h(\delta) = \frac{1}{2}\eta_1\eta_2\kappa_{fcd}\min\left\{\frac{\eta_2}{\kappa_{bhm}}, 1\right\}\delta^2 := C_2\delta^2$. *Then we have*

$$\phi(X_k) - \phi(X_{k+1}) \geq \begin{cases} h(\Delta_k) - \mathcal{E}_k - \mathcal{E}_k^+ - r & \text{if } \Theta_k = 1 \\ -\mathcal{E}_k - \mathcal{E}_k^+ - r & \text{if } \Theta_k' = 1 \\ 0 & \text{if } \Theta_k' = 0. \end{cases}$$

Lemma (total progress)

For any positive integer T, *we have*

$$h(\tau\bar{\Delta})\sum_{k=0}^{T-1}\Theta_k\Lambda_k' \leq \phi(x_0) - \hat{\phi} + \sum_{k=0}^{T-1}\Theta_k'\left(\mathcal{E}_k + \mathcal{E}_k^+ + r\right).$$

Lemma of Total Progress

Lemma (total progress)

For any positive integer T, we have

$$h(\tau\bar{\Delta})\sum_{k=0}^{T-1}\Theta_k\Lambda'_k \leq \phi(x_0) - \hat{\phi} + \sum_{k=0}^{T-1}\Theta'_k\left(\mathcal{E}_k + \mathcal{E}_k^+ + r\right).$$

$$h(\tau\bar{\Delta}) = C_2\left(\tau C_1 \min\{\|\nabla\phi(X_k)\| :\ k = 1, 2, \ldots T-1\}\right)^2$$

Classification of Iterations

<table>
<tr><th></th><th colspan="3">$I_k = 1, J_k = 1$</th><th colspan="3">$I_k = 1, J_k = 0$</th><th colspan="3">$I_k = 0, J_k = 1$</th><th colspan="3">$I_k = 0, J_k = 0$</th></tr>
<tr><th></th><th>★</th><th>✓</th><th>✗</th><th>★</th><th>✓</th><th>✗</th><th>★</th><th>✓</th><th>✗</th><th>★</th><th>✓</th><th>✗</th></tr>
<tr><td>$\Delta_k \in (\bar{\Delta}, \infty)$</td><td>1</td><td>4</td><td>5</td><td>6</td><td>9</td><td>11</td><td>13</td><td>16</td><td>18</td><td>20</td><td>23</td><td>25</td></tr>
<tr><td>$\Delta_k \in (\tau\bar{\Delta}, \bar{\Delta}]$</td><td>2</td><td colspan="2" rowspan="2"></td><td>7</td><td rowspan="2">10</td><td rowspan="2">12</td><td>14</td><td rowspan="2">17</td><td rowspan="2">19</td><td>21</td><td rowspan="2">24</td><td rowspan="2">26</td></tr>
<tr><td>$\Delta_k \in (0, \tau\bar{\Delta}]$</td><td>3</td><td>8</td><td>15</td><td>22</td></tr>
</table>

Lemma (**sufficient condition for successful step**)

If $I_k J_k = 1$ *and* $\Lambda_k = 0$ *then* $\Theta_k = 1$.

Classification of Iterations

	$I_k = 1, J_k = 1$			$I_k = 1, J_k = 0$			$I_k = 0, J_k = 1$			$I_k = 0, J_k = 0$		
	★	✓	✗	★	✓	✗	★	✓	✗	★	✓	✗
$\Delta_k \in (\bar{\Delta}, \infty)$	1	4	5	6	9	11	13	16	18	20	23	25
$\Delta_k \in (\tau\bar{\Delta}, \bar{\Delta}]$	2			7	10	12	14	17	19	21	24	26
$\Delta_k \in (0, \tau\bar{\Delta}]$	3			8			15			22		

Lemma (**total progress**)

For any positive integer T, we have

$$h(\tau\bar{\Delta}) \sum_{k=0}^{T-1} \Theta_k \Lambda'_k \leq \phi(x_0) - \hat{\phi} + \sum_{k=0}^{T-1} \Theta'_k \left(\mathcal{E}_k + \mathcal{E}_k^+ + r\right).$$

Lemma of Total Progress

Since $\{(\mathcal{E}_k, \mathcal{E}_k^+)\}$ are subexponential and independent

Lemma (**total loss**)

For any $t \geq 0$*,*

$$\mathbb{P}\left\{\sum_{k=0}^{T-1}\left(\mathcal{E}_k + \mathcal{E}_k^+ + r\right) \geq T(4/a + 2\epsilon_f + r) + t\right\} \leq \exp\left(-\frac{a}{4}t\right).$$

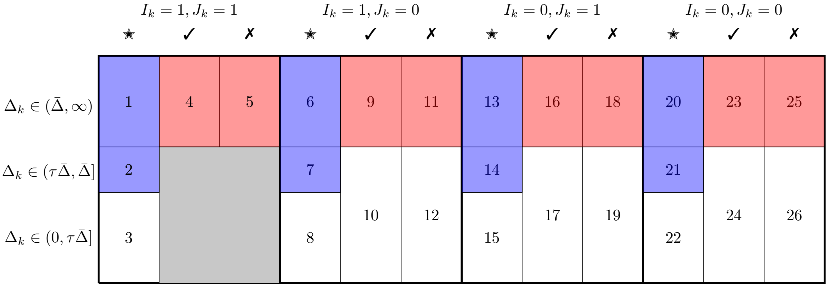
Lemma (**total progress**)

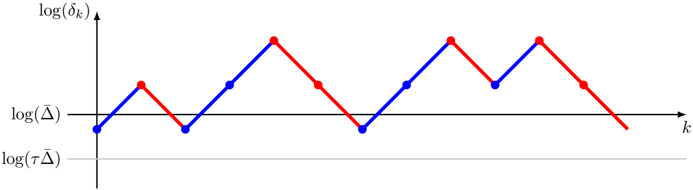
For any positive integer T*, we have*

$$\begin{aligned} h(\tau\bar{\Delta})\sum_{k=0}^{T-1}\Theta_k\Lambda_k' &\leq \phi(x_0) - \hat{\phi} + \sum_{k=0}^{T-1}\Theta_k'\left(\mathcal{E}_k + \mathcal{E}_k^+ + r\right) \\ &< \phi(x_0) - \hat{\phi} + T(4/a + 2\epsilon_f + r) + t \end{aligned}$$

with probability at lease $1 - \exp\left(-\frac{a}{4}t\right)$*.*

Ups and Downs of the Radius

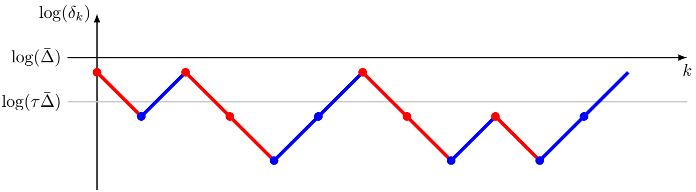




$$\sum_{k=0}^{T-1} (1 - \Theta_k)\Lambda_k \leq \sum_{k=0}^{T-1} \Theta_k \Lambda'_k + \log_\tau \left(\frac{\bar{\Delta}}{\delta_0} \right) + 1$$

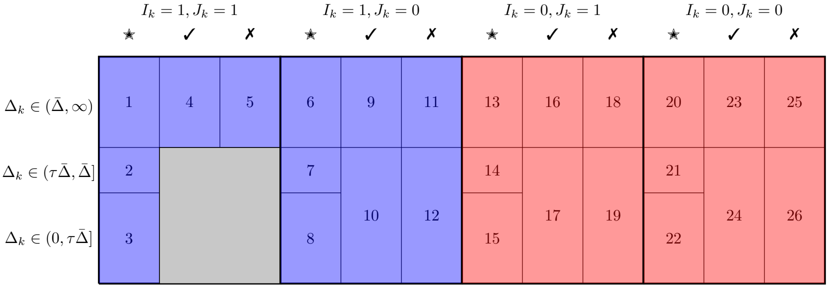
Downs and Ups of the Radius

	$I_k = 1, J_k = 1$			$I_k = 1, J_k = 0$			$I_k = 0, J_k = 1$			$I_k = 0, J_k = 0$		
	★	✓	✗	★	✓	✗	★	✓	✗	★	✓	✗
$\Delta_k \in (\bar{\Delta}, \infty)$	1	4	5	6	9	11	13	16	18	20	23	25
$\Delta_k \in (\tau\bar{\Delta}, \bar{\Delta}]$	2			7	10	12	14	17	19	21	24	26
$\Delta_k \in (0, \tau\bar{\Delta}]$	3			8			15			22		



$$\sum_{k=0}^{T-1} \Theta_k (1 - \Lambda'_k) \leq \sum_{k=0}^{T-1} (1 - \Theta_k)(1 - \Lambda_k)$$

Iterations with Sufficiently Accurate Gradient Estimate

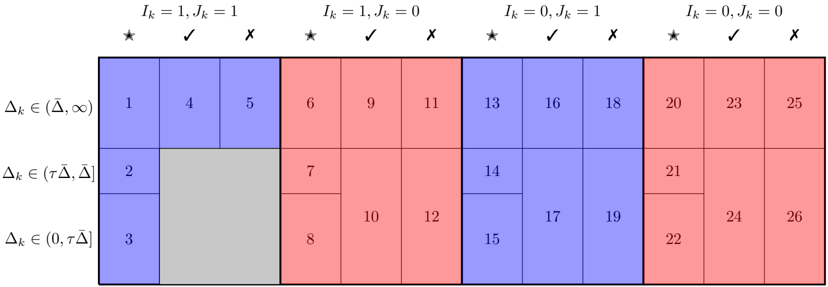


Lemma

Assume $\mathbb{P}\{I_k = 1 \mid \mathcal{F}_{k-1}\} \geq p_1$ *holds. By Azuma-Hoeffding inequality, for any positive integer* T *and any* $\hat{p}_1 \in [0, p_1]$ *we have*

$$\mathbb{P}\left\{\sum_{k=0}^{T-1} I_k > \hat{p}_1 T\right\} \geq 1 - \exp\left(-\frac{(1-\hat{p}_1/p_1)^2}{2} T\right).$$

Iterations with Sufficiently Accurate Function Evaluation



Lemma

Assume both $\mathbb{P}\{\mathcal{E}_k > t\}$ *and* $\mathbb{P}\{\mathcal{E}_k^+ > t\}$ *are* $\leq \exp(a(\epsilon_f - t))$. *Let* $p_0 = 1 - 2\exp(a[\epsilon_f - r/2])$. *For any positive integer* T *and any* $\hat{p}_0 \in [0, p_0]$, *we have*

$$\mathbb{P}\left\{\sum_{k=0}^{T-1} J_k > \hat{p}_0 T\right\} \geq 1 - \exp\left(-2(p_0 - \hat{p}_0)^2 T\right).$$

Analysis

$$\sum_{k=0}^{T-1}(1-\Theta_k)\Lambda_k \leq \sum_{k=0}^{T-1}\Theta_k\Lambda'_k + \log_\tau\left(\frac{\bar{\Delta}}{\delta_0}\right) + 1$$

$$\sum_{k=0}^{T-1}\Theta_k(1-\Lambda'_k) \leq \sum_{k=0}^{T-1}(1-\Theta_k)(1-\Lambda_k)$$

$$\mathbb{P}\left\{\sum_{k=0}^{T-1} I_k > \hat{p}_1 T\right\} \geq 1 - \exp\left(-\frac{(1-\hat{p}_1/p_1)^2}{2}T\right)$$

$$\mathbb{P}\left\{\sum_{k=0}^{T-1} J_k > \hat{p}_0 T\right\} \geq 1 - \exp\left(-2(p_0-\hat{p}_0)^2 T\right)$$

$$\Downarrow$$

$$\mathbb{P}\left\{\sum_{k=0}^{T-1}\Theta_k\Lambda'_k > \left(\hat{p}_0+\hat{p}_1-\frac{3}{2}\right)T - \frac{1}{2}\log_\tau\left(\frac{\bar{\Delta}}{\delta_0}\right) - \frac{1}{2}\right\}$$
$$\geq 1 - \exp\left(-\frac{(1-\hat{p}_1/p_1)^2}{2}T\right) - \exp\left(-2(p_0-\hat{p}_0)^2 T\right)$$

Main Result

Theorem

Let assumptions hold. Given any $\epsilon > \sqrt{\frac{4\epsilon_f+8/a+2r}{C_2\tau^2C_1^2(2p_0+2p_1-3)}}$, *we have*

$$\mathbb{P}\left\{\min\{\|\nabla\phi(X_k)\|:\ 0\le k\le T-1\}\le\epsilon\right\}\ge$$

$$1-\exp\left(-\frac{(1-\hat{p}_1/p_1)^2}{2}T\right)-\exp\left(-2(p_0-\hat{p}_0)^2T\right)-\exp\left(-\frac{a}{4}t\right)$$

for any $\hat{p}_0$ *and* $\hat{p}_1$ *such that* $\hat{p}_0+\hat{p}_1\in\left(\frac{3}{2}+\frac{2\epsilon_f+4/a+r}{C_2(\tau C_1\epsilon)^2}, p_0+p_1\right]$, *any* $t\ge 0$, *and any*

$$T\ge\left(\hat{p}_0+\hat{p}_1-\frac{3}{2}-\frac{2\epsilon_f+4/a+r}{C_1(\tau C_2\epsilon)^2}\right)^{-1}\left[\frac{\phi(x_0)-\hat{\phi}+t}{C_1(\tau C_2\epsilon)^2}+\frac{1}{2}\log_\tau\left(\frac{C_2\epsilon}{\delta_0}\right)+\frac{1}{2}\right].$$

Main Result

Theorem

Let assumptions hold. Given any $\epsilon > \sqrt{\frac{4\epsilon_f + 8/a + 2r}{C_2\tau^2 C_1^2(2p_0+2p_1-3)}}$, *we have*

$$\mathbb{P}\left\{\min\{\|\nabla\phi(X_k)\| : \ 0 \le k \le T-1\} \le \epsilon\right\} \ge 1 - \exp\left(-\frac{(1-\hat{p}_1/p_1)^2}{2}T\right) - \exp\left(-2(p_0-\hat{p}_0)^2 T\right) - \exp\left(-\frac{a}{4}t\right)$$

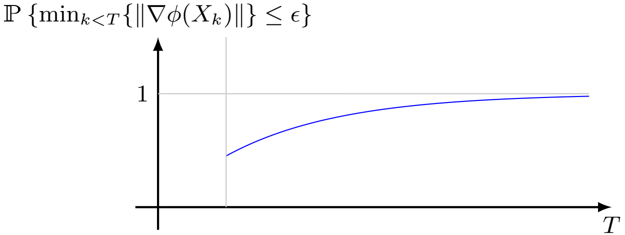
for any $\hat{p}_0$ *and* $\hat{p}_1$ *such that* $\hat{p}_0 + \hat{p}_1 \in \left(\frac{3}{2} + \frac{2\epsilon_f + 4/a + r}{C_2(\tau C_1\epsilon)^2}, p_0 + p_1\right]$, *any* $t \ge 0$, *and any*

$$T \ge \left(\hat{p}_0 + \hat{p}_1 - \frac{3}{2} - \frac{2\epsilon_f + 4/a + r}{C_1(\tau C_2 \epsilon)^2}\right)^{-1}\left[\frac{\phi(x_0) - \hat{\phi} + t}{C_1(\tau C_2\epsilon)^2} + \frac{1}{2}\log_\tau\left(\frac{C_2\epsilon}{\delta_0}\right) + \frac{1}{2}\right].$$

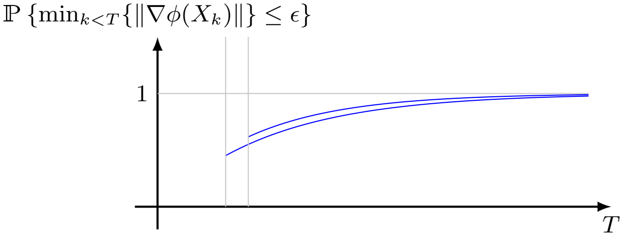
Every parameter is either from the algorithm or an assumption expect $\hat{p}_0, \hat{p}_1$ and t. We need to optimize the final result over them.

In addition, recall $\rho_k = \frac{f(x_k) - f(x_k+s_k) + r}{m(x_k) - m(x_k+s_k)}$ from the algorithm. What value should we use for r?

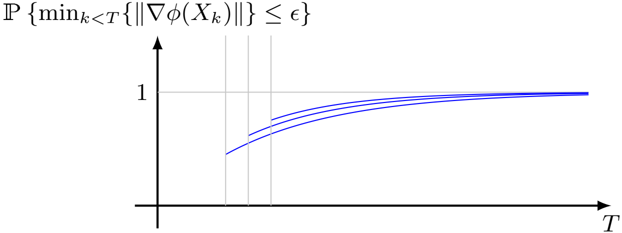
Main Result



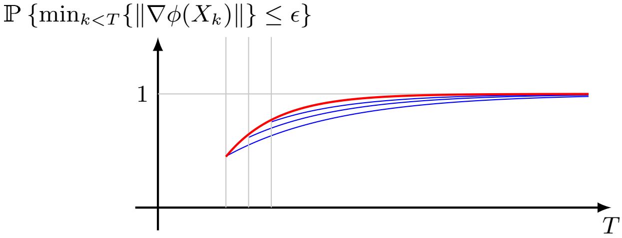
Main Result



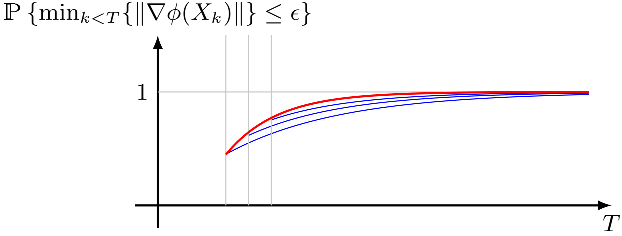
Main Result



Main Result

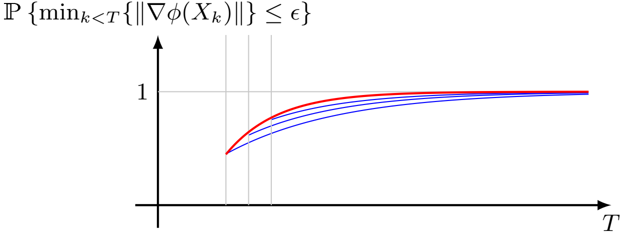


Main Result



Unfortunately the optimal value for $\hat{p}_0, \hat{p}_1$ and t can not be expressed with basic mathematical operations.

Main Result



Unfortunately the optimal value for $\hat{p}_0, \hat{p}_1$ and t can not be expressed with basic mathematical operations.

Fortunately we have the optimal value

$$r = 2\epsilon_f + \frac{2}{a}\log\left(aC_1(\tau C_2\epsilon)^2\right).$$

Remarks

- We also have result where we assume the noise is bounded like $|e(x)| \leq \epsilon_f$ instead of being subexponential. The optimal value for $\hat{p}_1$ has closed form.
- We are working on the second-order convergence proof.
- We hoped to use the weaker assumption $\mathbb{P}\left\{e(X_k) - e(X_k^+) + r > 0\right\} \geq p_0$. It might be possible because the loss at any accepted iteration is bounded:

$$\phi(x_{k+1}) \leq \phi(x_k) + \langle \nabla\phi(x_k), s_k \rangle + \frac{L}{2}\|s_k\|^2 \leq \phi(x_k) + \|\nabla\phi(x_k)\|\delta_k + \frac{L}{2}\delta_k^2.$$

Remarks

- We also have result where we assume the noise is bounded like $|e(x)| \leq \epsilon_f$ instead of being subexponential. The optimal value for $\hat{p}_1$ has closed form.
- We are working on the second-order convergence proof.
- We hoped to use the weaker assumption $\mathbb{P}\left\{e(X_k) - e(X_k^+) + r > 0\right\} \geq p_0$. It might be possible because the loss at any accepted iteration is bounded:

$$\phi(x_{k+1}) \leq \phi(x_k) + \langle \nabla\phi(x_k), s_k \rangle + \frac{L}{2}\|s_k\|^2 \leq \phi(x_k) + \|\nabla\phi(x_k)\|\delta_k + \frac{L}{2}\delta_k^2.$$

Thank you!