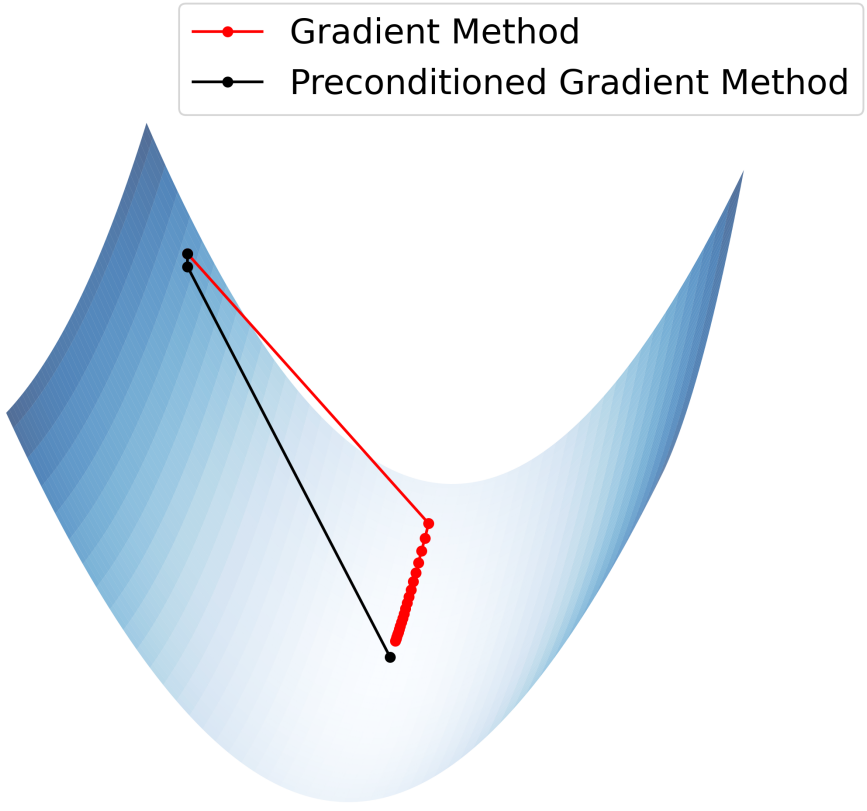
Rank-one approximation of the Hessian

short presentation

Andrei Semenov

MLO Group Meeting, 04.02.2025

Intro: Preconditioning



$$x^+ = x - \gamma \hat{D}^{-1} \nabla f(x)$$

Intro: Examples

- Hutchinson's: $D = \mathbb{E}\left[z \odot \nabla^2 f(x) z\right]$, $z \sim \text{Radamacher}$
- Fisher Information Matrix: $D = \mathbb{E}\left[\nabla \log p(b \mid a)(\nabla \log p(b \mid a)^\top)\right]$
- Adam's preconditioner: $D^2 = \text{diag}\left(\nabla f(x) \odot \nabla f(x)\right)$
- Hessian: $\gamma = 1$, $D = \nabla^2 f(x) \rightarrow$ Newton's method

We study Rank-1 preconditioners

- **Uniform Direction** preconditioner:

$$D = n\nabla^2 f(x) uu^\top,$$

u sampled form a unit sphere $\mathbb{S}^{n-1} \subset \mathbb{R}^n$

- **Fisher Information Matrix**:

$$D = \mathbb{E}\left[-\nabla \log p(b \mid a)\right] = \mathbb{E}\left[\nabla \log p(b \mid a) \nabla \log p(b \mid a)^\top\right],$$

a – inputs, b – labels.

Uniform Direction preconditioner

$$D = n\nabla^2 f(x) uu^\top$$

Pros ?

Uniform Direction preconditioner

$$D = n\nabla^2 f(x) uu^\top$$

Pros ?

$$\mathbb{E}[D] = \nabla^2 f(x)$$

Uniform Direction preconditioner

$$D = n\nabla^2 f(x) uu^\top$$

Pros ?

$$\mathbb{E}[D] = \nabla^2 f(x)$$

Cons ?

Uniform Direction preconditioner

$$D = n\nabla^2 f(x) uu^\top$$

Pros ?

$$\mathbb{E}[D] = \nabla^2 f(x)$$

Cons ?

- non-symmetric

Uniform Direction preconditioner

$$D = n\nabla^2 f(x) uu^\top$$

Pros ?

$$\mathbb{E}[D] = \nabla^2 f(x)$$

Cons ?

- non-symmetric
- not PSD

Uniform Direction preconditioner

We can do symmetrization

$$S = \frac{n}{2}\left(\nabla^2 f(x)uu^\top + \left(\nabla^2 f(x)uu^\top\right)^\top\right).$$

Uniform Direction preconditioner

We can do symmetrization

$$S = \frac{n}{2}\left(\nabla^2 f(x) u u^\top + \left(\nabla^2 f(x) u u^\top\right)^\top\right).$$

We can add a regularizer to D

$$D = S + \delta I.$$

Uniform Direction preconditioner

We can do symmetrization

$$S = \frac{n}{2}\left(\nabla^2 f(x) uu^\top + \left(\nabla^2 f(x) uu^\top\right)^\top\right).$$

We can add a regularizer to D

$$D = S + \delta I.$$

But here we lose the unbiasedness

Uniform Direction preconditioner

How far are we from the true Hessian?

Uniform Direction preconditioner

How far are we from the true Hessian?

Proposition

$$\mathbb{E}\left[\|D - \nabla^2 f(x)\|_F^2\right] = \lambda_{\max}^2 n\,(n-1) + n\delta^2$$

Uniform Direction preconditioner

$$x^+ = x - (D + \alpha I)^{-1} \nabla f(x),$$

$$D = S + \delta I.$$

Uniform Direction preconditioner

$$x^+ = x - (D + \alpha I)^{-1} \nabla f(x),$$

$$D = S + \delta I.$$

$$\sigma := \sup_{x \in \mathbb{R}^n} \mathbb{E}\left[\|D - \nabla^2 f(x)\|_F \right]$$

Uniform Direction preconditioner

$$x^+ = x - (D + \alpha I)^{-1} \nabla f(x),$$

$$D = S + \delta I.$$

$$\sigma := \sup_{x \in \mathbb{R}^n} \mathbb{E}\left[\| D - \nabla^2 f(x) \|_F \right]$$

$$\sigma \leq \sqrt{\mathbb{E}\left[\| D - \nabla^2 f(x) \|_F^2 \right]} = \sqrt{\lambda_{\max}^2 n (n-1) + n\delta^2}$$

Uniform Direction preconditioner

$$x^+ = x - (D + \alpha I)^{-1} \nabla f(x),$$

$$D = S + \delta I.$$

$$\sigma := \sup_{x \in \mathbb{R}^n} \mathbb{E}\left[\|D - \nabla^2 f(x)\|_F \right]$$

$$\sigma \leq \sqrt{\mathbb{E}\left[\|D - \nabla^2 f(x)\|_F^2 \right]} = \sqrt{\lambda_{\max}^2 n\,(n-1) + n\delta^2}$$

$$\delta = 0 \Rightarrow \sigma \leq \lambda_{\max} \sqrt{n(n-1)}$$

Convergence. Gradient Method with Uniform Direction

We consider

$$\min_{x \in \mathbb{R}^n} f(x) \tag{1}$$

And we use the following assumptions

Convergence. Gradient Method with Uniform Direction

We consider

$$\min_{x \in \mathbb{R}^n} f(x) \tag{1}$$

And we use the following assumptions

Assumptions

We assume that the Hessian is Lipschitz continuous, with parameter $M > 0$:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq M\|x - y\| \quad \forall x, y \in \mathbb{R}^n.$$

Convergence. Gradient Method with Uniform Direction

We consider

$$\min_{x \in \mathbb{R}^n} f(x) \tag{1}$$

And we use the following assumptions

Assumptions

We assume that the Hessian is Lipschitz continuous, with parameter $M > 0$:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq M\|x - y\| \quad \forall x, y \in \mathbb{R}^n.$$

Note, that $f(x)$ can be non-convex.

Convergence. Gradient Method with Uniform Direction

We consider

$$\min_{x \in \mathbb{R}^n} f(x) \tag{1}$$

And we use the following assumptions

Assumptions

We assume that the Hessian is Lipschitz continuous, with parameter $M > 0$:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq M\|x - y\| \quad \forall x, y \in \mathbb{R}^n.$$

Note, that $f(x)$ can be non-convex.

$$x_{k+1} = x_k - (D_k + \alpha_k I)^{-1} \nabla f(x_k) \tag{2}$$

Convergence. Gradient Method with Uniform Direction

Lemma: Convergence of Gradient Method with Uniform Direction preconditioner

Let f has a M-Lipschitz Hessian and bounded parameter σ. Consider Algorithm 2 with

$$\alpha_k = \sqrt{\frac{M\|\nabla f(x_k)\|}{2}} + \sigma,$$
$$\sigma = \sqrt{\lambda_{\max}^2 n(n-1) + n\delta^2}$$

for $k \geq 1$. Then, for any $\varepsilon > 0$, it is enough to do

$$K = \left\lceil 8(f(x_0) - f^*) \cdot \left(\sqrt{\frac{M}{2}} \frac{1}{\varepsilon^{3/2}} + \frac{\sigma}{\varepsilon^2} \right) + 2\log \frac{\|\nabla f(x_0)\|}{2} \right\rceil, \quad (3)$$

steps to ensure $\min_{1 \leq k \leq K} \|\nabla f(x_k)\| \leq \varepsilon$.

Experiments for Gradient Method with Uniform Direction

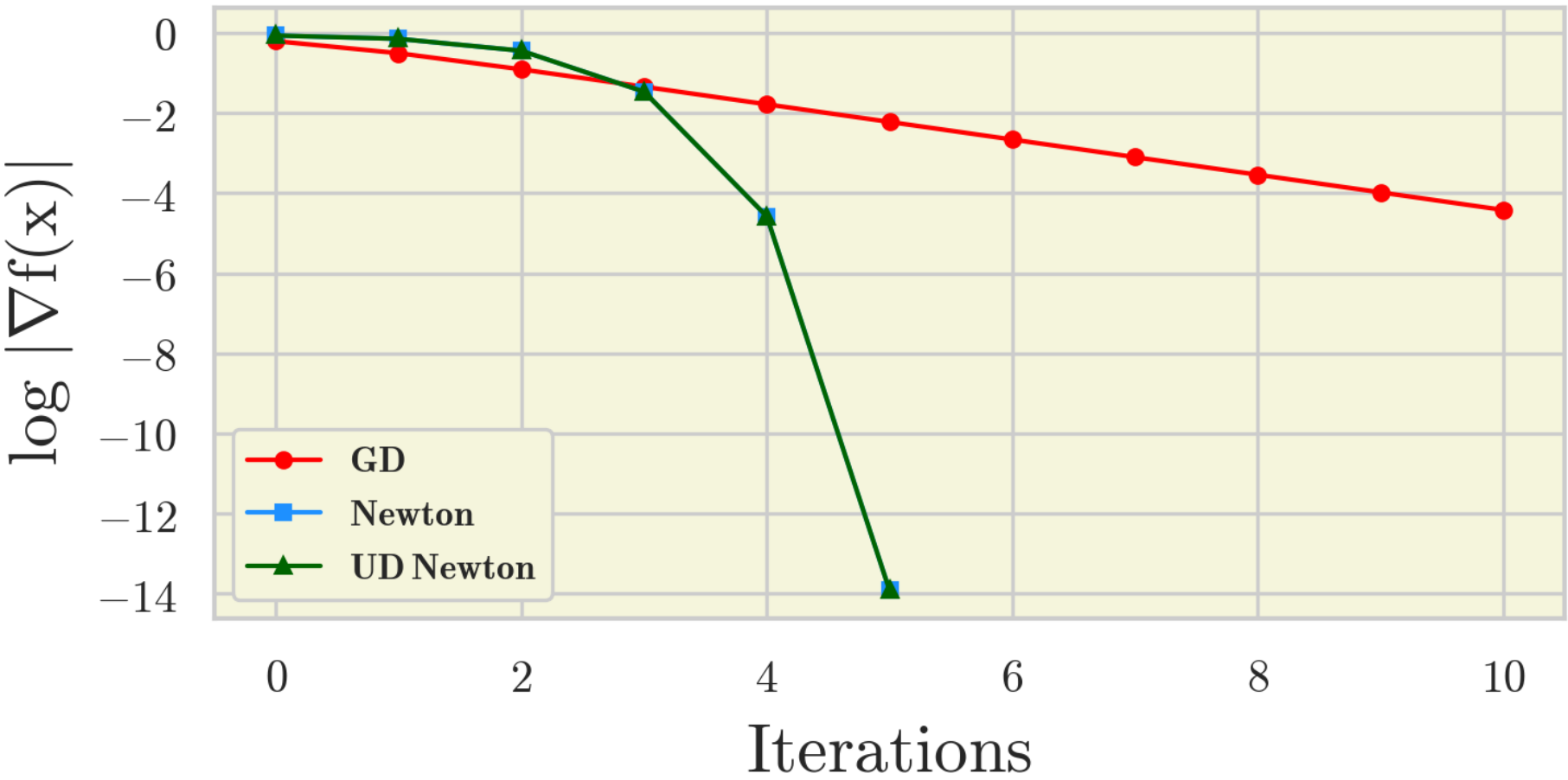


Figure: 1D function, $\nabla^2 f(x) \succ 0$

Experiments for Gradient Method with Uniform Direction

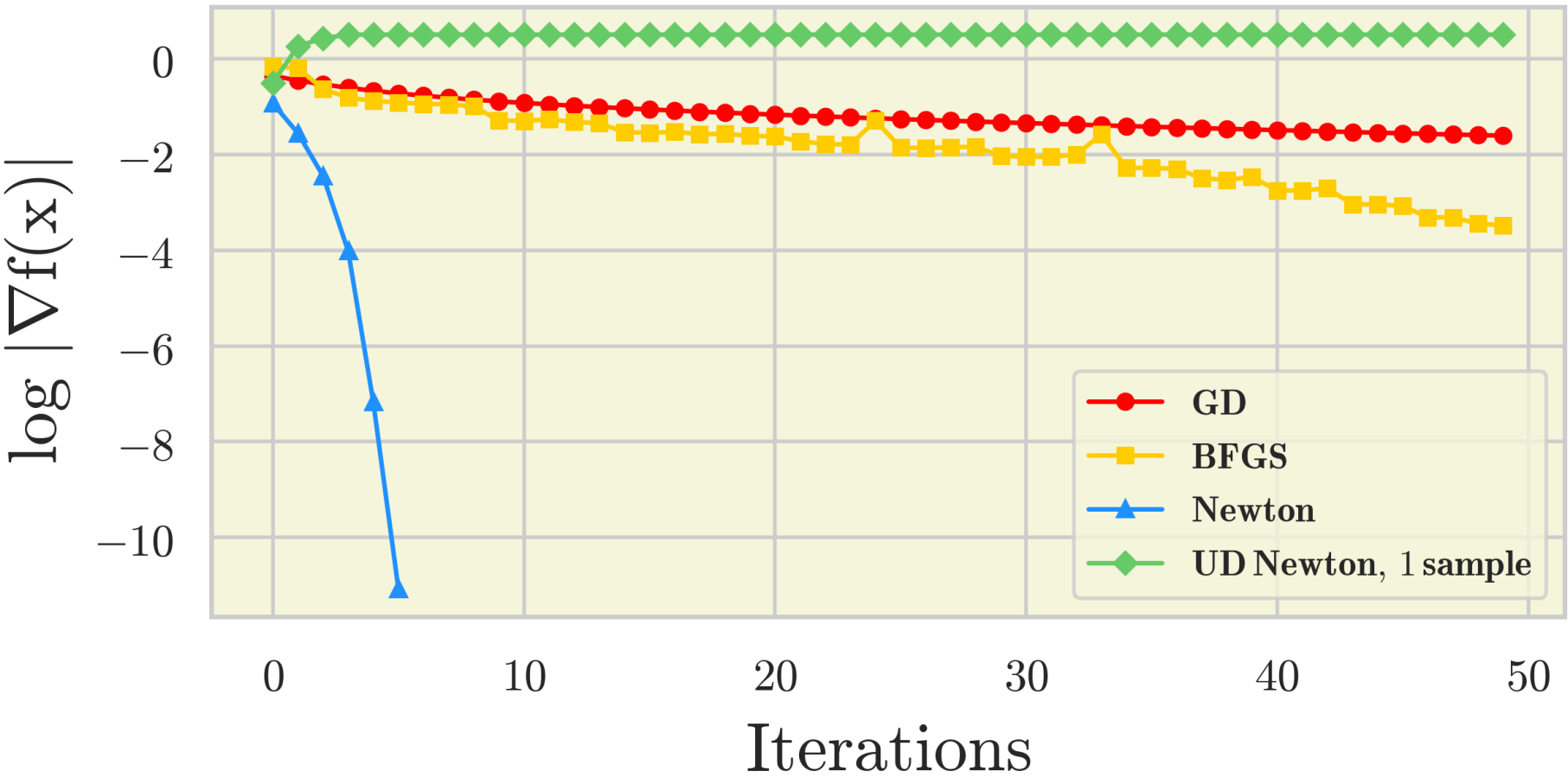


Figure: Mushrooms, $\nabla^2 f(x) \in \mathbb{R}^{112\times 112}$

Experiments for Gradient Method with Uniform Direction

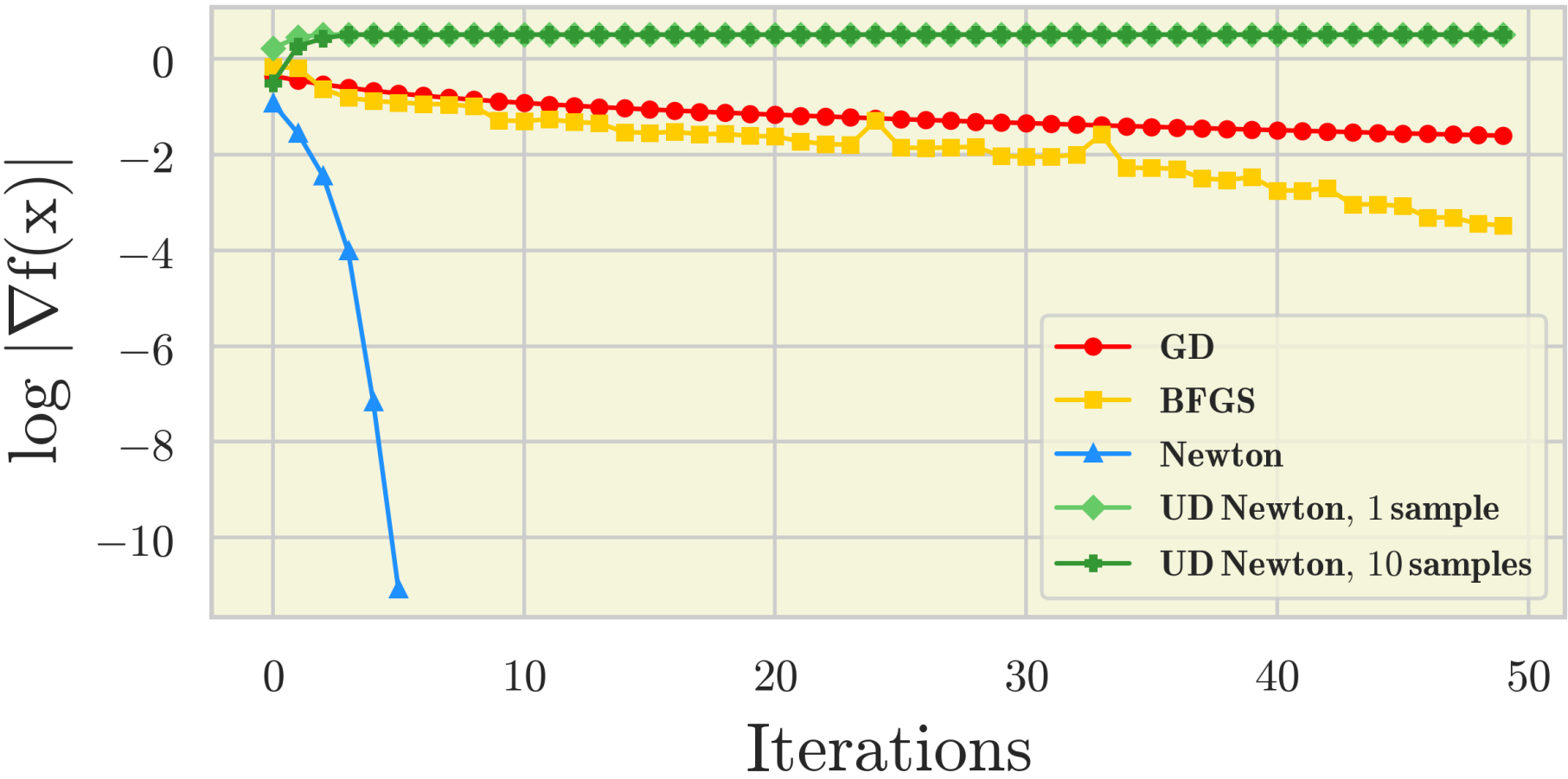


Figure: Mushrooms, $\nabla^2 f(x) \in \mathbb{R}^{112\times 112}$

Experiments for Gradient Method with Uniform Direction

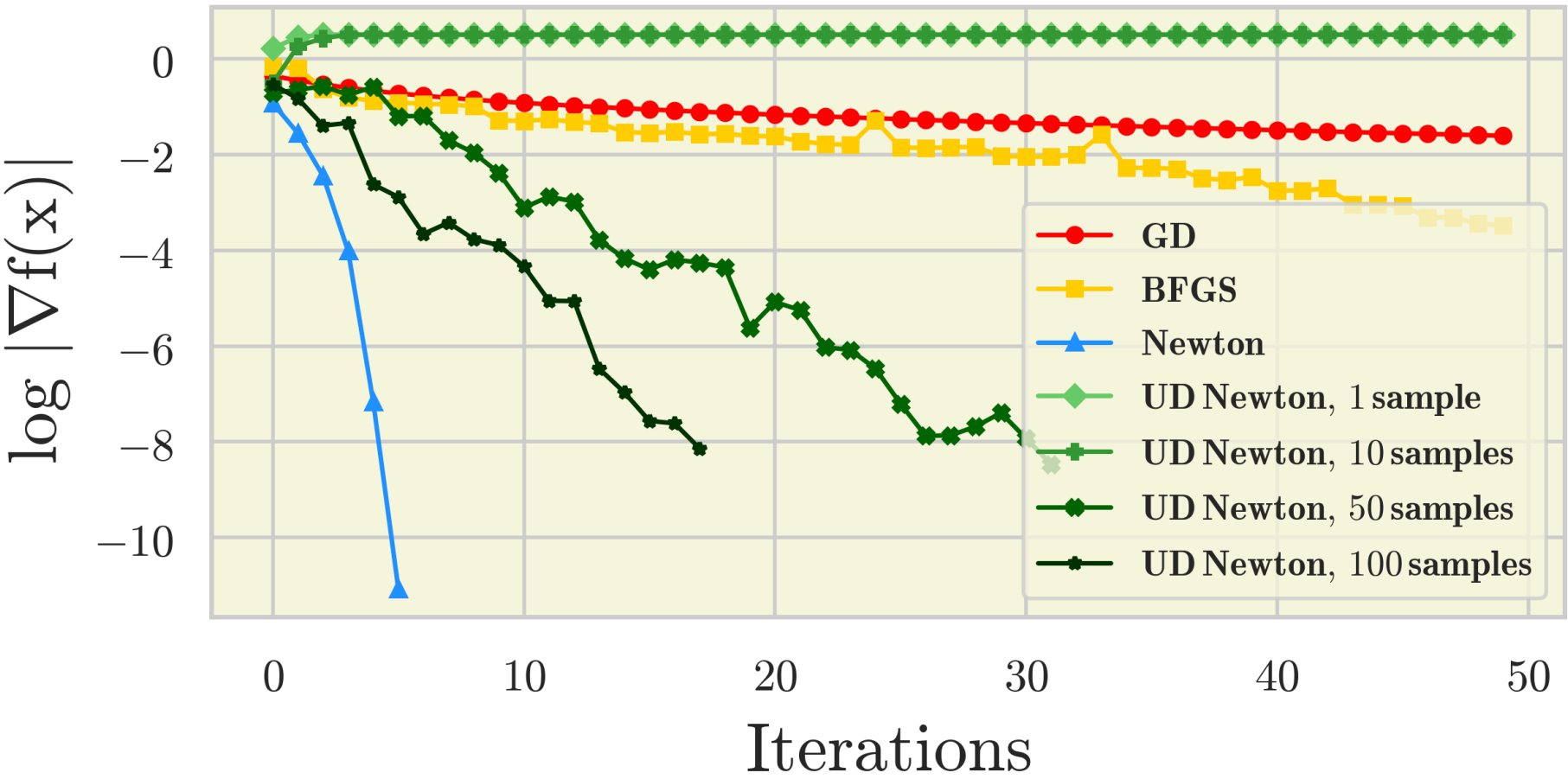


Figure: Mushrooms, $\nabla^2 f(x) \in \mathbb{R}^{112\times 112}$

Experiments for Gradient Method with Uniform Direction

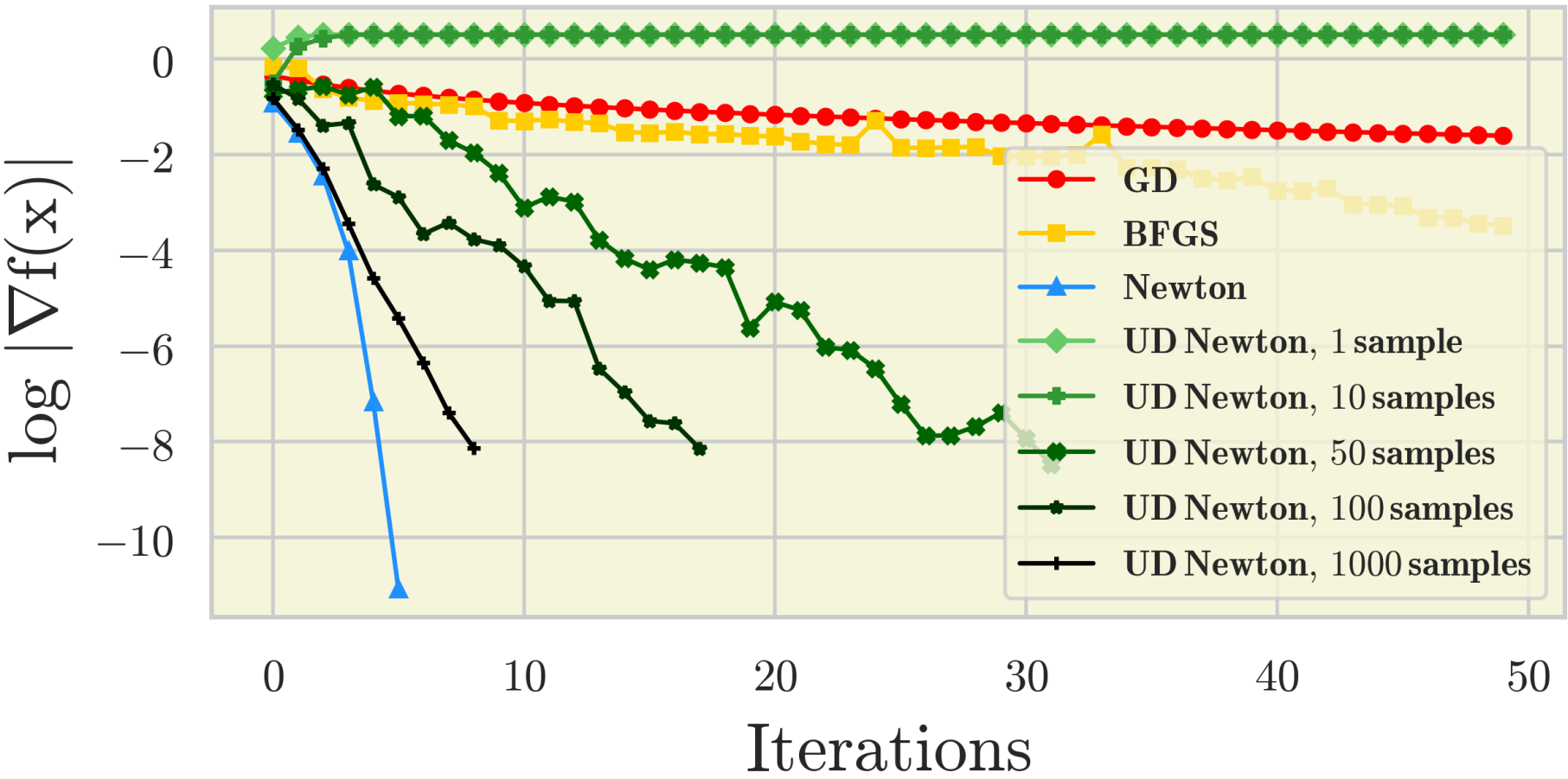


Figure: Mushrooms, $\nabla^2 f(x) \in \mathbb{R}^{112\times 112}$

Fisher preconditioner

$$x_{k+1} = x_k - (D_k + \alpha_k I)^{-1} \nabla f(x_k), \tag{4}$$

$$D_k = \mathbb{E}\left[\nabla \log p_{x_k}(b \mid a)(\nabla \log p_{x_k}(b \mid a)^\top)\right].$$

Fisher preconditioner

$$x_{k+1} = x_k - (D_k + \alpha_k I)^{-1} \nabla f(x_k), \tag{4}$$
$$D_k = \mathbb{E}\left[\nabla \log p_{x_k}(b \mid a)(\nabla \log p_{x_k}(b \mid a)^\top)\right].$$

Informal: Let φ be L-smooth "inner function", which represents the model's outputs (logits) just before the loss f.

Fisher preconditioner

$$x_{k+1} = x_k - (D_k + \alpha_k I)^{-1} \nabla f(x_k), \tag{4}$$
$$D_k = \mathbb{E}\left[\nabla \log p_{x_k}(b \mid a)(\nabla \log p_{x_k}(b \mid a)^\top)\right].$$

Informal: Let φ be L-smooth "inner function", which represents the model's outputs (logits) just before the loss f.

Lemma: credits to Frederik

$$\|\nabla^2 f(x) - F(x)\|_2^2 \leq L \sum_{s=1}^{S} \|\nabla_\varphi \log p_x(b_s \mid \varphi(a_s, x))\|_1.$$

Here S is the size of the entire dataset.

Convergence. Gradient Method with Fisher Information Matrix

Assumptions

We assume that the Hessian is Lipschitz continuous, with parameter $M > 0$:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq M\|x - y\| \quad \forall x, y \in \mathbb{R}^n.$$

In addition we assume the convexity of loss f:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall x, y \in \mathbb{R}^n.$$

Convergence. Gradient Method with Fisher Information Matrix

$$\sigma := \sup_{x \in \mathbb{R}^n} \|\nabla^2 f(x) - D(x)\|$$

Convergence. Gradient Method with Fisher Information Matrix

$$\sigma := \sup_{x \in \mathbb{R}^n} \|\nabla^2 f(x) - D(x)\|$$

We consider two special cases, for now

Convergence. Gradient Method with Fisher Information Matrix

$$\sigma := \sup_{x \in \mathbb{R}^n} \|\nabla^2 f(x) - D(x)\|$$

We consider two special cases, for now

- nonlinear least squares:

$$f(x) = \frac{1}{2} \sum_{s=1}^{S} (\varphi(a_s, x) - b_s)^2 ,$$

Convergence. Gradient Method with Fisher Information Matrix

$$\sigma := \sup_{x \in \mathbb{R}^n} \|\nabla^2 f(x) - D(x)\|$$

We consider two special cases, for now

- nonlinear least squares:

$$f(x) = \frac{1}{2} \sum_{s=1}^{S} (\varphi(a_s, x) - b_s)^2 ,$$

- softmax:

$$f(x) = \sum_{s=1}^{S} \frac{\exp \varphi_{b_s}(a_s)}{\sum_{l=1}^{S} \exp \varphi_l(a_s)}$$

Convergence. Gradient Method with Fisher Information Matrix

$$\sigma := \sup_{x \in \mathbb{R}^n} \|\nabla^2 f(x) - D(x)\|$$

We consider two special cases, for now

- nonlinear least squares:

$$f(x) = \frac{1}{2} \sum_{s=1}^{S} (\varphi(a_s, x) - b_s)^2 ,$$

- softmax:

$$f(x) = \sum_{s=1}^{S} \frac{\exp \varphi_{b_s}(a_s)}{\sum_{l=1}^{S} \exp \varphi_l(a_s)}$$

Convergence. Gradient Method with Fisher Information Matrix. Least Squares

Let us start with least squares.

Convergence. Gradient Method with Fisher Information Matrix. Least Squares

Let us start with least squares.

$$\sigma \leq L \sum_{s=1}^{S} \mid \varphi(a_s, x) - b_s \mid \leq L\sqrt{2S}\sqrt{f(x)},$$

Convergence. Gradient Method with Fisher Information Matrix. Least Squares

Let us start with least squares.

$$\sigma \leq L \sum_{s=1}^{S} \mid \varphi(a_s, x) - b_s \mid \leq L\sqrt{2S}\sqrt{f(x)},$$

$$\sigma \leq L\sqrt{2S}\left(\sqrt{f(x) - f^*} + \sqrt{f^*}\right)$$

Convergence. Gradient Method with Fisher Information Matrix. Least Squares

Lemma: Convergence of Gradient Method with Fisher Information Matrix. Least Squares

Consider Algorithm 4 with

$$\alpha_k = \sqrt{\frac{M\|\nabla f(x_k)\|}{2}} + \sigma(x_k), \quad \sigma(x_k) \leq L\sqrt{2S}\left(\sqrt{f(x_k) - f^*} + \sqrt{f^*}\right),$$

for $k \geq 1$. Then, for any $\varepsilon > 0$, it is enough to do

$$K = \left\lceil 8\left(f(x_0) - f^*\right) \cdot \left(\left(\sqrt{\frac{M}{2}} + L\sqrt{2SR}\right)\varepsilon^{-3/2} + L\sqrt{2Sf^*}\varepsilon^{-2}\right) + 2\log\frac{\|\nabla f(x_0)\|}{\varepsilon}\right\rceil,$$

steps to ensure $\min_{1\leq k\leq K}\|\nabla f(x_k)\| \leq \varepsilon$.

Convergence. Gradient Method with Fisher Information Matrix. Least Squares

Now, we consider the softmax case.

Convergence. Gradient Method with Fisher Information Matrix. Least Squares

Now, we consider the softmax case.

$$\sigma \leq \left\| \nabla_{\varphi} \log \frac{e^{\varphi_{b_s}}}{\sum_j e^{\varphi_j}} \right\|_1 = \left\| \nabla_{\varphi} \left(\varphi_{b_s} - \log \left(\sum_j e^{\varphi_j} \right) \right) \right\|_1$$
$$= 2 \left(1 - \frac{e^{\varphi_{b_s}}}{\sum_j e^{\varphi_j}} \right) \leq -2 \log \frac{e^{\varphi_{b_n}}}{\sum_j e^{\varphi_j}},$$

Convergence. Gradient Method with Fisher Information Matrix. Least Squares

Now, we consider the softmax case.

$$\sigma \leq \left\| \nabla_{\varphi} \log \frac{e^{\varphi_{b_s}}}{\sum_j e^{\varphi_j}} \right\|_1 = \left\| \nabla_{\varphi} \left(\varphi_{b_s} - \log \left(\sum_j e^{\varphi_j} \right) \right) \right\|_1$$
$$= 2 \left(1 - \frac{e^{\varphi_{b_s}}}{\sum_j e^{\varphi_j}} \right) \leq -2 \log \frac{e^{\varphi_{b_n}}}{\sum_j e^{\varphi_j}},$$

$$\sigma \leq 2f(x)$$

Convergence. Gradient Method with Fisher Information Matrix. Softmax

Lemma: Convergence of Gradient Method with Fisher Information Matrix. Softmax

Consider Algorithm 4 with

$$\alpha_k = \sqrt{\frac{M\|\nabla f(x_k)\|}{2}} + \sigma(x_k), \quad \sigma(x_k) \leq 2\left(f(x_k) - f(x_{k+1})\right) + 2f(x_{k+1}),$$

for $k \geq 1$. Then, for any $\varepsilon > 0$, it is enough to do

$$K = \left\lceil 8\left(f(x_0) - f^*\right) \cdot \left(2R\varepsilon^{-1} + \sqrt{\frac{M}{2}}\varepsilon^{-3/2} + 2f^*\varepsilon^{-2}\right) + 2\log\frac{\|\nabla f(x_0)\|}{\varepsilon}\right\rceil,$$

steps to ensure $\min_{1 \leq k \leq K} \|\nabla f(x_k)\| \leq \varepsilon$.

The End

Thanks!